Supplementary Material for "Impact of predicted protein-truncating genetic variants on the human transcriptome"

Manuel A. Rivas, Matti Pirinen, Donald F. Conrad, Monkol Lek,
Emily K. Tsang, Konrad J. Karczewski, Julian B. Maller, Kimberly R. Kukurba,
David DeLuca, Menachem Fromer, Pedro G. Ferreira, Kevin S. Smith
Rui Zhang, Fengmei Zhao, Eric Banks, Ryan Poplin,
Douglas Ruderfer, Shaun M. Purcell, Taru Tukiainen,
Peter D. Stenson, David N. Cooper, Katharine H. Huang, Timothy J. Sullivan,
Jared Nedzel, the GTEx Consortium, the Geuvadis Consortium,
Carlos D. Bustamante, Jin Billy Li, Mark J. Daly, Roderic Guigo, Peter Donnelly,
Kristin Ardlie, Michael Sammeth, Emmanouil Dermitzakis,
Mark I. McCarthy, Stephen B. Montgomery,
Tuuli Lappalainen, and Daniel G. MacArthur

Contents

List of Tables

List of Figures

S1 Study Design

An integrated data set was generated to study the impact of predicted protein-truncating genetic variants (PTVs) on the human transcriptome. We combined exome and multi-tissue transcriptome data from 173 individuals (up to 30 tissues per individual) with genome and lymphoblastoid cell line (LCL) transcriptome sequencing data from 462 individuals (genotypes for 41 individuals were imputed). We used mRNA quantifications (gene, splice junction, exon and intron) along with allelic ratio data herein referred to as allele-specific expression (ASE) data to:

1. study the transcriptional properties of PTV-containing transcripts,
2. gain insights into nonsense-mediated decay (NMD),
3. gain insights into the impact of large structural gene deletions on gene expression and assess evidence for dosage compensation for heterozygous PTVs, and
4. gain insights into the transcriptional impact of variants proximal to splice junctions.

We focused on nonsense single nucleotide variants (SNVs), frameshift indels, splice-disrupting SNVs, and large deletions.

S2 Genotype data

S2.1 Geuvadis data set

S2.1.1 SNVs and indels

We used the genotype data from the Phase 1 release of the 1000 Genomes project (1KG) for 462 individuals (genotypes for 41 individuals were imputed) included in Lappalainen et al. 2013 and available at `http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/genotypes/` (*5*, *9*).

S2.1.2 Large deletions

The large deletion calls for GEUVADIS samples were extracted from the official Phase 1 data release of the 1000 Genomes project.

All files used to identify the breakpoints, genotypes and validation status of these calls can be downloaded (as of March 11, 2014) from:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120813_phase1_large_del_submitted_to_DGVA/`

Much of the data may also be obtained from the Database of Genomic Variants archive:

`http://www.ebi.ac.uk/dgva/`

Validation. Three experimental approaches were used by the Structural Variation (SV) subgroup of the 1000 Genomes project to validate their deletion calls. We summarize the validation methods here, more complete details can be found in the supplemental material for the Phase 1 paper (*9*).

- **IRS.** An Illumina Omni 2.5M genotyping chip was run on all samples. Probe intensity data from this chip was used to perform *in silico* validation and estimate false discovery rate of deletion call. A rank-sum test was used to compute a *P* value for each deletion call, where a call consists of a combination of a deletion site (chromosome, start, end) plus a list of samples called as carriers of the deletion (samples that are either het or homozygous deleted). To evaluate a call,

the intensities for each probe falling within the deletion site were ranked separately and then the ranks were re-ranked across all probes within the called deletion site, using random order to break ties. A Mann-Whitney-Wilcoxon (MWW) test was used to compute a *P* value for each call based on the combined rankings. When confidence intervals on the boundaries were supplied for a deletion call, only probes within the innermost confidence interval were used to evaluate the call.

- **Polymerase chain reaction (PCR).** SV callsets were created by multiple groups involved in the 1000 Genomes project. To enable the calculation of false discovery rates (FDRs) for independent SV callsets, the SV subgroup randomly picked 96 loci from each deletion callset for subsequent PCR validation experiments. The randomization was carried out by randomly picking, without replacement, from the entire list of generated calls for each SV discovery callset. Duplicate primers between different callsets were removed, yielding 91-96 loci tested per callset. PCR validation experiments of deletions were designed using a spanning primer strategy where both primers hybridize to regions flanking the predicted SV. PCRs resulted in either a band size corresponding to the reference allele, or a shorter amplicon corresponding to the reference allele band size reduced by the inferred SV size. In addition to the putative deletion carrier, each PCR was run along with three controls: NA12892 genomic DNA, a pool of five DNAs, corresponding to four human samples (HG00407 + HG00689 + NA18507 + NA19314 (Coriell) and a chimpanzee sample EB176 (JC) (HPA Culture Collections).

- **Array comparative genomic hybridization (CGH).** A custom Agilent 2x1M CGH Microarray was designed against the union of SV calls made in a 25-sample subset. Each SV locus was genotyped in this set of 25 samples. A locus was considered "validated" if the number of samples with a validated deletion call was greater than the number of samples in which a deletion call could not be validated.

There are 22,531 deletions in the Phase 1 release. We considered a deletion to be validated if it passed at least one of the three validation experiments, and it did not fail PCR or CGH validation. This selection criteria results in 8,021 validated deletions from samples in the GEUVADIS collection.

The 1KG deletion calls were produced by a large combination of first-generation tools for copy number variation (CNV) discovery from next-sequencing data. Due to the complexities of merging call sets, a locus could be considered validated yet have significant breakpoint misestimation or genotyping error. To ensure the highest possibly quality of the data, we manually curated $1,425$ PTV deletion events by inspecting the LRR and BAFs from the Omni2.5 data. These represent all events that passed some basic filters (biallelic, rare, partial or complete PTV) and which had > 4 probes spanning the deletion region on Omni2.5. This produced a short, conservative list of 145 deletion calls with extremely high confidence for a total of 59 unique PTV deletions. Importantly for our PTV analyses, a variant was only considered validated if the original reported breakpoints of the deletion were consistent with the breakpoints apparent from the array data. Ultimately we restricted our definition of large deletions to just those that remove the entire coding DNA sequence (CDS) of a gene, due to technical concerns about the interplay between structural variation and mapping accuracy producing a list of 33 large deletions pre manual curation and a list of 25 large deletions post manual curation.

S2.2 GTEx data set

S2.2.1 Exome sequencing

We performed whole exome sequencing on blood DNA samples from 180 GTEx pilot phase donors at the Broad Institute's Genomics Platform, using Agilent Sure-Select Human All Exon v2.0, 44Mb baited target, and the Broad in-solution hybrid selection process. For input DNA we used $>$ 250 ng

of DNA, at $>$ 2 ng/ul. Our exome-sequencing pipeline included sample plating, library preparation (2-plexing of samples per hybridization), hybrid capture, sequencing (76bp paired reads), sample identification QC check, and data storage. Our hybrid selection libraries cover $> 80\%$ of targets at 20x and a mean target coverage $>$ 80x. The exome sequencing data was de-multiplexed and each samples sequence data were aggregated into a single Picard BAM file.

S2.2.2 SNVs and indels

Exome sequencing data was processed through a pipeline based on Picard from `http://picard.sourceforge.net/`, using base quality score recalibration and local realignment at known indels (these algorithms were originally presented in (*28*) describing the GATK sofware, but has since been migrated to Picard). We used the BWA aligner from `http://bio-bwa.sourceforge.net` for mapping reads to the human genome build 37 (hg19) (*29*). GATKs UnifiedGenotyper package was then used for SNV calling jointly across all 180 samples (*28*, *30*). GATKs HaplotypeCaller (v2.8) was used for indel calling across all 180 samples (*31*). We applied default filters to SNV and indel calls using GATK's Variant Quality Score Recalibration (VQSR) approach and generated a VCF file (*32*).

S2.2.3 Large deletions

Large deletion genotyping from Array data All GTEx pilot samples were run on two Illumina Human microarray platforms: the Omni 5M and the Exome array. The Log_2R ratios from each array were quantile normalized to a single reference distribution, and then these transformed values were merged into a single data set for each sample, what we refer to as the "SuperArray". Instead of doing *ab initio* large deletion discovery and genotyping, we attempted to genotype all known common deletion loci that contained at least one SuperArray probe, using methods implemented in CNVtools (*33*). We targeted 2,593 deletions with frequency greater than 1% in either the CEU or YRI samples of 1000 Genomes pilot 1, and we were able to genotype 488 sites using the SuperArray. Following QC we were left with high quality genotypes for 30 common PTV deletions.

Large deletion calling from exome sequencing We used the XHMM software to detect copy number variation in whole-exome sequencing data from the GTEx project in 180 individuals (*34*). The XHMM exome sequencing CNV discovery and genotyping pipeline (*34*, *35*) was run on these samples to detect exon-level copy number variation and assign CNV quality metrics.

XHMM output files were converted to PLINK format for QC purposes. Stringent call-level QC was performed by removing all sex chromosome CNV and low-quality XHMM calls (XHMM SQ < 60).

As detailed in the standard XHMM protocol, coverage was calculated and averaged over each target using GATK DepthOfCoverage and XHMM. Mean sample coverage ranged between 40x and 80x for most samples (fig. S1a).

Approximately 4000 targets were found to have little or no coverage in most samples (fig. S1b), and thus removed; otherwise, per-target coverage necessarily averaged at 60x, with a long tail with some targetes at 200-300x coverage. This is typical of exome-sequencing experiments and thus necessitates normalization across targets before CNV can be called (*34*). Using XHMM, normalization was performed in a principal component analysis framework, where the 7 largest principal components were automatically normalized out (fig. S1c); as is typical, these largest components have a qualitatively different character than subsequent components and are correlated with combinations of GC content, and sample and target read depths (fig. S1d), which reflect both genomic properties of the exons being targeted as well as the varying sample preparations such as batch effects. After normalization, most targets have a relatively low variance among samples, but we removed the remaining

outlier targets for which a large spread among the sample read depths still existed, possibly indicating failed normalization.

We excluded all multi-allelic events from analyses and subsequently removed 8 outlier samples (those with no CNV or greater than > 50 CNV calls. In the default QC pipeline it is recommended to remove CNVs larger than 5Mb. We ultimately decided not to use this filter given the independent technology (array) available for confirmation).

The resulting set of CNVs calls was then distilled into a single CNV map, where we consider two CNV calls with greater than 50% reciprocal overlap to be alleles of the same locus. This CNV map consisted of 1413 unique loci, split into 399 sites of deletion, 626 sites of duplication, and 370 sites where both deletions and duplications were observed.

Validation Common deletions genotyped by the Omni 5M array were considered validated if they overlapped a known common large deletion locus reported by either the 1000 Genomes project (*9*) or the Structural Variation Consortium (*36*).

The goal of our rare deletion validation analysis was to create a conservative set of heterozygous deletion calls that have strong experimental support. XHMM rare deletion calls were validated by comparison with probe intensity data from the combined Omni 5M and exome arrays that were on all GTEx samples (the "SuperArray"). The SuperArray contained at least one probe within the predicted breakpoints of 539 of our 636 XHMM deletion calls. We summarized the copy number of each putative deletion region as the average probe Log_2R ratio within the region. We observed a clear validation signal when we inspected both the rank (fig. S2a) and the mean Log_2R ratio of putative deletions (fig. S2b). We defined our final, validated deletion callset as deletions with XHMM genotype quality score greater than 70, SuperArray rank less than 5, and Log_2R ratio greater than -3 (the latter to avoid homozygous deletions). This produced a validation callset of 70 rare deletions. Ultimately we restricted our definition of large deletions to just those that remove the entire CDS of a gene (as described in earlier section), due to technical concerns about the interplay between structural variation and mapping accuracy producing a list of 8 large deletions (after the removal of "2" large deletions identified to overlap with a 20Mb mosaic event).

S2.3 Variant Annotation

S2.3.1 SNVs and indels

Annotation of SNVs and small indels was performed using a modified version of the Variant Effect Predictor (VEP v2.5; `http://useast.ensembl.org/info/docs/variation/vep/vep_script.html`) tool from Ensembl and Gencode v12 annotation. As shown in a previous study, the choice of transcripts and software may have a large effect on variant annotation (*22*) and we choose the GENCODE v12 gene models to maintain consistency across all the analyzed data sets. The tool was modified to produce custom annotation tags (table S1). The additional annotation tags were applied to variants that were annotated as STOP_GAINED, SPLICE_DONOR_VARIANT, SPLICE_ACCEPTOR_VARIANT, or FRAME_SHIFT and flagged if any filters failed. A PTV is predicted as high confidence (HC) if there is at least one transcript that passes all filters described in MacArthur et al (2012), (*4*). Nonsense-mediated decay (NMD) prediction was performed using the PLINK/SEQ v0.09 toolkit from `https://bitbucket.org/statgen/plinkseq/` and described in detail in section S7.3.

S2.3.2 Large Deletion Annotation

To maintain consistency with previous work we used the same pipeline described in MacArthur et al. to annotate large deletions (*4*). Briefly, this pipeline considers 6 ways in which a deletion can result in severe disruption of a gene: deletion of first exon, deletion that results in a shift in the reading

frame, deletion of the full coding DNA sequence (CDS), deletion of a part of the CDS, deletion of a splice site, and deletion of the start codon. We annotated all validated deletion calls using GENCODE v12 transcripts. A deletion was considered to produce "complete" loss-of-function of a gene if all transcripts of that gene were annotated with one of the 6 possible PTV classifications. The precise location of the breakpoints for array and sequencing based deletion calls is often uncertain; array-based breakpoints typically having uncertainty in the range of several kb, while sequencing breakpoints may be exact or associated with anywhere from 1bp - 1kb of uncertainty. Our annotation pipeline considers the precision of the breakpoint estimates for each deletion, and uses these to produce a conservative PTV annotation - if any possible definition of the deletion breakpoints would exclude/reverse a finding of PTV derived from the estimated breakpoints, then the deletion will not be annotated as a PTV. Ultimately we restricted our definition of PTV large deletions to just those that remove the entire CDS of a gene.

S3 RNA sequencing

S3.1 Geuvadis

As described in Lappalainen et al. 2013 and 't Hoen et al. 2013, (*5*, *6*) LCLs were collected from Coriell Cell Repositories (GBR, FIN, TSI) or originally from Coriell but grown at the University of Geneva (CEU, YRI). The live cultures were shipped to ECACC (European Collection of Cell Cultures) as live cultures. In ECACC, the cell lines were cultured, then split to produce $8\times$ cell banks of the samples, and finally shipped to Geneva. RNA was extracted in Geneva and assessed for RNA quality and quantity.

RNA sequencing of the Epstein-Barr virus (EBV) growth-transformed lymphoblastoid cell lines (LCLs) was performed in multiple European sequencing facilities on the Illumina HiSeq2000 platform with 75bp paired-end sequencing with fragment size of approximately 280 bp using the Illumina TruSeq library construction protocol. This is a non-strand specific polyA+ selected library.

Furthermore, as indicated in Lappalainen et al. 2013 (*5*) each lab submitted one demultiplexed fastq file. Reads failing Illumina quality filtering were removed. mRNA read mapping was conducted with the GEM aligner (*37*).

S3.2 GTEx

RNA sequencing of the tissues collected in the pilot phase of the GTEx project was performed using the Illumina TruSeq library construction protocol. This is a non-strand specific polyA+ selected library. The sequencing produced 76-bp paired end reads. Further detail on the samples, read alignment, post-processing, tissue and sample collection is described in the main GTEx analysis manuscript (*8*).

We used the quantifications for the nine tissues in GTEx with the highest number of subjects:

1. Adipose Subcutaneous (ADPSBQ, $n = 94$);
2. Artery Tibial (ARTTBL, $n = 112$);
3. Heart Left Ventricle (HRTLV, $n = 83$);
4. Lung (LUNG, $n = 119$);
5. Muscle Skeletal (MSCLSK, $n = 138$);
6. Nerve Tibial (NERVET, $n = 88$);

7. Skin Sun Exposed (SKINS, $n = 96$);

8. Thyroid (THYROID, $n = 105$); and

9. Blood (WHLBLD, $n = 155$).

S4 mRNA quantifications

S4.1 Gene, exon, intron, and splice junction quantifications

From the aligned RNA-seq read data from both GTEx and Geuvadis, we quantified several features of transcripts based on the GENCODE v12 gene annotation: genes, exons, introns, and splice junctions. Gene and exon quantifications for the GTEx and the Geuvadis data sets were generated using very similar methods briefly described in this section.

The data files of these quantifications are available in

- Geuvadis: `http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/`,
- GTEx: `http://www.broadinstitute.org/gtex/datasets`.

Briefly, for exon quantifications, overlapping exons in the annotations were merged, and reads overlapping these regions were counted. Gene quantifications for GTEx were the sum of the exons per gene converted to RPKM, and in Geuvadis the sum of transcript RPKM quantifications was used. For intron quantifications, we counted the number of reads that overlap regions that are exclusively spanned by introns. Splice junctions were quantified by counting the number of reads mapping to splice-junctions with split-mappings with consecutive read positions mapping to the last (donor) and first (acceptor) exonic positions in the annotated splice-junctions. Only properly paired reads were used in the quantifications, and in all but exonic quantifications only reads complying with the annotated exon-intron structure were included in the analysis.

S4.2 Normalization of quantifications

In the analysis of this paper, unless otherwise specified, for normalization of the quantifications we used the methods in Lappalainen et al. 2013 (*5*) also for GTEx data, so that we were able to combine GEUVADIS and GTEx data. In general, these methods are very similar to those used in the main GTEx analysis manuscript (*8*). The supplementary material of Lappalainen et al. (*5*) provides additional details of the quantification and normalization methods used in this project, and they are briefly summarized below: All read count quantifications were first normalized by sequencing depth by dividing them by the total number of mapped reads per sample. For exon, intron, and junction quantifications we scaled them to the median number of mapped reads (85M in GTEx). We then split the data to tissues and all subsequent analysis was done separately for each of the tissues with > 80 samples.

We filtered elements (genes/exons/junctions etc) to keep only those with > 0 expression $> 50\%$ of the individuals, except for introns where this filtering step was omitted (since we are searching for rare *inclusions* of introns). We further removed technical variation using PEER (*38*) similarly to Lappalainen et al. (*5*): we ran PEER for 20,000 quantification units (e.g. exons or genes) with 10 factors ($K = 10$), adding the mean to the model. Covariates from this analysis were regressed out from all the quantifications, and the mean was added to the residuals. These quantifications were further transformed to standard normal distribution for the splice disruption analysis.

S5 Transcriptional properties of PTV-containing transcripts

Each human carries at least 100 PTVs even after careful filtering for sequencing and annotation errors, but the majority of these variants in an individual appear to be common in the population and reside in genes that are likely tolerant to dramatic disruption (*4*). One putative mechanism underlying such tolerance is that PTVs often do not affect all transcripts of a gene (*4*), and tissue-specific expression of the different transcripts may potentially affect the penetrance of the PTV. We studied the overall gene expression properties of PTV containing genes and the expression of splice junctions containing splice disrupting variants.

S5.1 Expression properties

We identified PTVs in the 462 individuals in the Lappalainen et al. data set (*5*) and the 180 exome sequencing data set in the GTEx project. We partitioned the PTVs into categories by minor allele frequency (MAF): a) common (MAF $\geq .05$; 1,607 genes), b) low frequency ($.01 <$ MAF $< .05$; 864 genes), and c) rare (MAF $\leq .01$; 5,096 genes). We compared these sets to the set of protein-coding genes (13,372 genes) where no PTVs were observed (fig. S3-7).

Comparison of tissue-wide gene expression profile for PTV containing groups of genes We compared the distribution of median gene expression values for PTV containing genes across tissues using the MWW one-sided (`wilcox.test` in R, `alternative="less"`) test (fig. S4).

Comparative analysis of ubiquitous expression We used the same grouping of genes to assess how the different groups of genes are expressed across all tissues. We compared the proportion of genes with median gene expression value above a $\log$(RPKM) cutoff for "rare", "low frequency", "common", and "no PTV" protein-coding genes (fig. S5).

Comparative analysis of tissue-specific expression Then, we studied the tissue-specific expression properties. We transformed the median expression values for gene i and tissue j as follows: Let $\mathbf{A}$ be an expression matrix with I protein-coding genes (rows) and J tissues (columns) with values in RPKM. Let $\mathbf{B}$ be the matrix of tissue specificities for protein-coding gene i and tissue j with the transformation

$$B_{ij} = \frac{A_{ij}}{\|A_i\|}, \qquad \text{(Eq. 1)}$$

where $\|A_i\|$ is the ℓ^2 norm of the gene expression vector for protein-coding gene i.

We compared the proportion of protein-coding genes above a tissue-specificity measure cutoff for "rare", "low frequency", "common", and "no PTV" protein-coding genes (fig. S6).

S5.2 Splice junction usage comparison

We compared the usage of splice-junctions - where we identify splice-disrupting variants - by analyzing splice-junction measurements of individuals with homozygote reference genotypes.

We used the number of reads spanning (normalized by the number of reads per sample) the splice junction and divided by the gene RPKM to get a relative splice junction abundance measure (Relative SJ Abundance, fig. S7). Variation in the Human Gene Mutation Database (HGMD) version hgmd-2012.4 was used in the analysis (*2*).

S6 Allele-specific analysis

S6.1 Allele-specific analysis of SNVs

Allele-specific expression analysis was based on allelic counts in the RNA-seq reads of heterozygous sites genotyped from DNA, within each individual. In this analysis of GTEx data, we included only heterozygous genotypes with genotype quality ≥ 60 (Phred-scale) for SNVs and ≥ 95 for indels. For the Geuvadis data set the maximum genotype quality in the VCF is 50 - we used this as the genotype quality threshold.

Additionally, we excluded sites that are susceptible to allelic mapping bias: 1) sites with 50bp mappability < 1 based on the UCSC mapability track, implying that the 50bp flanking region of the site is non-unique in the genome, and 2) simulated RNA-seq reads overlapping the site show $> 5\%$ difference in the mapping of reads that carry the reference or alternate allele (*39*). We only used uniquely mapping reads (mapping quality > 150), NM ≥ 6, and sites with base quality > 10. We included sites in the analyses with at least 8 reads in the heterozygote individual.

S6.2 Allele-specific analysis of indels

S6.2.1 Background

Frameshift insertions and deletions (indels) are believed to be an important source of loss-of-function variation, and are known to contribute to several disease associations (*4*, *40*). However, proper characterization of these variants has been difficult. The first problem is that calling small indels from genome and exome sequencing data has been challenging, with even the best algorithms having high error rates especially for low frequency variants (*12*).

Furthermore, while transcriptome effects and NMD caused by indels can in principle be analyzed by ASE in heterozygote individuals, this has not been done before due to difficulty in obtaining reliable RNA-seq read counts of the reference and alternate alleles. This is due to two challenges:

1. If RNA-seq reads are aligned to the reference genome, any read carrying alternative indel alleles will have several mismatches, and often fail to map in the correct location. This leads to a higher allelic mapping bias than in SNVs (*39*);

2. Even in cases when the reads map to the correct locus of the reference genome, there are often small errors in the local alignment, with mappers often failing to decipher the exact location and type of the alternative allele.

One possible solution to the problem is realignment of the RNA-seq data to full personalized genomes. However, in large data sets such as GTEx and GEUVADIS, the creation of 1270 maternal and paternal reference genomes, the alignment itself, and further analysis and storage of terabytes of BAM files would have been computationally very demanding. This is further complicated by the fact that indels alter the genomic coordinates.

S6.2.2 Local personalized mapping for ASE analysis

Thus, we developed a new approach to address the challenges in indel ASE analysis in a computationally feasible manner. In addition to the outline of the approach here, the scripts and documentation of detailed parameters are available in `http://tllab.org/data-software/`. Our method is based on alignment to local reference sequences that have been modified to contain both the reference and alternative alleles. The workflow is shown in fig. S8.

Briefly, for each variant we extract the flanking ($\pm$ 100bp) reference genomic sequence, and modify it to build an alternative allele reference index. For the analysis of this study, we chose to use

only the GTEx data with GATK HaplotypeCaller genotype calls from exome sequencing data. Our analysis showed that while the GATK UnifiedGenotyper produced good-quality genotypes for SNVs that were used for the other analyses of this paper, for indels the highest concordance with RNA-seq data was obtained with GATK HaplotypeCaller (fig. S9). We performed ASE analysis using the local modified reference mapping for the following categories of autosomal variants:

1. 500 randomly selected synonymous SNVs from exome sequence data as a control for SNVs;
2. 1845 nonsense predicted protein-truncating SNVs from exome sequence data;
3. 1540 inframe indels ($\leq$6bp) from exome sequence data as a control for indels;
4. 1628 frameshift predicted protein-truncating indels ($\leq$6bp) from exome sequence data; and
5. 130 frameshift predicted protein-truncating indels from exome-chip data to confirm that the results are not biased by genotype calling from exome sequence data.

We mapped all RNA-seq reads from 1634 samples from 173 subjects separately against the two reference indices, one with the reference and one with the alternative alleles of all the variants. The mapping was done with BWA, mapping the reads as single-end. These BAM files were then processed jointly to count the reference and alternative allele reads in each locus, based on reads that map perfectly to only one of the two reference indices. This, as well as the latter filtering step (see below) makes this method conservative. This is sufficient for our analysis of genome-wide phenomena, but for a well-powered analysis of any site of interest will require further methods development.

An important caveat of this method is that as the alignment is not done against the entire genome, some reads that originate from another locus may incorrectly appear as uniquely mapping to the analyzed loci. To avoid this, we further filtered the results to remove variants where the flanking 36bp region is not unique in the genome, as indicated by the mapability track of the UCSC Genome Browser.

S6.2.3 Results

Having allele counts for each sample, we further filtered the data to only have sites with ≥ 8 reads. To evaluate the performance of this method, we compared the allele ratios obtained with this approach to those from the standard ASE analysis for the 210 SNVs with ≥ 8 reads in both analyses. The correlation of the ratios is high (Pearson correlation $= 0.80$, $P < 2.2 \times 10^{-16}$), and the lack of bias towards either direction suggests that the deviations are driven by random loss of some reads in localized mapping (fig. S10). This renders strong support for the accuracy of alignment both in this and in the traditional ASE approach.

Next, we investigated concordance of genotype calls and RNA-sequencing data, knowing that indel genotyping can have a higher error rate. The allele ratios for all REF/REF, REF/ALT, and ALT/ALT genotypes are shown in fig. S11. As expected (and observed before (*5*)), some heterozygous genotypes are monoallelic, making it difficult to distinguish in individual cases whether this is driven by true ASE, e.g. by NMD, or genotyping error with homozygous genotypes appearing as heterozygous. This is a much bigger problem for indels, especially deletions, and thus in all the ASE analysis of indels, in order to validate the heterozygous genotype, we use only sites with median allelic ratio across all the tissues $\leq .95$ and $\geq .05$. Altogether, these filters leave us with ASE data from 62,846 genotypes in 1929 variants.

To analyze the overall variant signal of ASE and NMD from these data, we first sampled one heterozygous individual from each variant in order to give equal weight to rare and common variants, and repeated the sampling 500 times. The results show a clear deviation towards the loss of the alternate allele in frameshift indels and nonsense SNVs compared to control set of indel and SNVs (fig. S12).

S6.3 Quantifying RNA allelic ratios by microfluidic multiplex PCR sequencing (mmPCR-seq)

S6.3.1 mmPCR-seq validation

We used mmPCR-seq (*11*) as an orthogonal technology to compare allelic ratios measured by RNA-seq at some PTV sites as well as control sites. This validation experiment included sites for this study as well as for two others studies. For this study, we selected 945 exonic sites from 9 individuals with large numbers of tissues (121; 8-13 per individual) and large quantity of RNA. We successfully designed multiplex primers for 682 of the 945 sites (table S3). For convenience, details about the methods for the mmPCR-seq validation are reproduced in the subsections S6.3.2 mmPCR primer design, S6.3.3 cDNA library construction and mmPCR preamplification, S6.3.4 Fluidigm mmPCR and barcoding, and S6.3.5 next-generation sequencing.

S6.3.2 mmPCR primer design

Primers were designed with the yamPCR program (*41*), which excises a template sequence, designs candidate primers on this sequence with a modified version of Primer3, uses BLAST to search for matches of candidate primers, and assembles a multiplex primer pool while ensuring that primers in the pool do not interact with each other or generate multiple amplicons. We used a version of this program adapted to design primers based on transcript cDNA sequences (*11*). This version finds a transcript that includes the site of interest based on provided gene annotations and uses the cDNA sequence around the site as the template for primer design. It uses a BLAST database with one representative transcript per gene with all the gene's exons. We modified the program to ensure that the sites were at least 2 bp from the 3' end of either primer and at most 70 bp from the 5' end of one of the primers. Furthermore, to prevent primer design over variant sites, we masked any site that was polymorphic in at least one of the nine individuals based on genotypes from exome sequencing data. Most primers were designed to have amplicons between 150 and 400 bp long, but we allowed longer amplicons for about 10% of sites which otherwise did not have primers designed for them. In total, 48 pools of up to 20 primer pairs were designed. A Perl implementation of the program is available at `http://montgomerylab.stanford.edu/resources.html`.

S6.3.3 cDNA library construction and mmPCR preamplification

121 GTEx samples were sent to Stanford under MTA #IC2013-1482. Five hundred nanograms of total RNA from subject-derived tissue (GTEx Consortium, The Broad Institute) were converted to cDNA using iScript reverse transcriptase (BioRad). Each cDNA library (150 ng) was preamplified using a primer pool (50 uM each) covering 818 individual sites (the 682 sites for this project and 135 additional sites not relevant to this study) with KAPA 2G (5x), Fast PCR Kit (KAPA Biosystems), and the following amplification protocol: 95C 10 min; (95°C 15 sec; 60°C 4 min) x 2; (95°C 15 sec; 72°C 4 min) x 13. Pre-amplified libraries were magnetic-bead purified (Agencourt AMPure XP) and the resulting library concentrations were spot-checked by NanoDrop before proceeding to the Fluidigm mmPCR step.

S6.3.4 Fluidigm mmPCR and barcoding

A maximum of 20-plex primers (50 uM each) were combined to form 48 different primer pools. The cDNA libraries (3.75 ul) for the 121 samples and 12 technical replicates, 8 inter-array and 4 intra-array, were combined with KAPA2G (5x) mix and Fluidigm 20x Access Array Loading Reagent. Three Fluidigm Access Arrays were primed, mixed, amplified, and harvested as described in Zhang et al. (*11*). Each Fluidigm Access Array held up to 48 samples. Harvested samples were prepared for sequencing with adapters containing individual barcodes by PCR amplification using KAPA2G (2x)

mix and the following amplification protocol: 95°C 5 min; (95°C 30 sec; 55°C 30 sec; 72°C 1 min) x3; (95°C 30 sec; 72°C 1 min) x10; 72°C 5 min; 12°C end. One-tenth of the volume of each library was checked for integrity on a 1.5% agarose gel. Three microliters of each amplified library were pooled and column purified of excess adapters (QIAquick PCR Purification Kit, Qiagen).

S6.3.5 Next-generation sequencing

Pooled mmPCR cDNA libraries from each Fluidigm Access Array were checked by Qubit (Life Technologies) for concentration, by Agilent Bioanalyzer for average size, and adjusted to 12 pM. The three pooled libraries were sequenced separately on the MiSeq desktop sequencer using 150 cycle V3 cartridges (Illumina) yielding 75 bp paired-end reads.

S6.3.6 mmPCR-seq data analysis

De-multiplexed reads were trimmed of adapters using cutadapt 1.4.1 (*42*). All reads shorter than 20 bp after trimming were removed. The reads were then mapped to the human reference (GRCH37) using STAR 2.3.0 (*43*) (–outFilterScoreMinOverLread 0 –outFilterMatchNminOverLread 0) and TopHat 1.4.1 (*44*) (–mate-inner-dist 300 –mate-std-dev 500). Counts of reference and alterative alleles over assayed sites were parsed from the output of samtools mpileup over the sites.

Comparison of aligners STAR mapped more reads than TopHat to the sites targeted by mmPCR-seq (fig. S13). The correlation of alternate allelic ratios for TopHat and STAR was highly comparable (figs. S14 and S15). Before quality control filtering, the Pearson correlation is 0.89 ($P < 2.2 \times 10^{-16}$). After quality control filtering (total depth count > 150, ref allele count > 5, and non-ref allele count > 5), the Pearson correlation is 0.95 ($P < 2.2 \times 10^{-16}$). Except for one variant that was excluded from downstream analysis (rs1138349), there is high correlation between TopHat and Star alignments. Therefore, with the exception of some direct comparisons to RNA-seq in the next section, all analyses using the mmPCR-seq data were performed using the STAR alignments.

Comparison of alternate allele ratios measured by RNA-seq and by mmPCR-seq The RNA-seq data were aligned with TopHat, so we used the mmPCR-seq TopHat alignments to control for mapping differences when comparing the alternate allele ratios. We compared heterozygous sites with at least 30 RNA-seq reads and that passed mmPCR-seq quality control filters described above. There was moderate correlation (Pearson correlation 0.52, $P = 4.4 \times 10^{-15}$) between the allelic ratios measured by these two technologies, but there were some sites for which the estimates of these ratios differed substantially (fig. S16). In addition, there was high correlation for nonsense SNVs (Pearson correlation 0.79, $P = 7.3 \times 10^{-14}$) between the allelic ratios measured by these two technologies.

We also compared the alternate allele ratio for heterozygous INDEL sites for which we have both mmPCR-seq and RNA-seq data and find significant correlation (Pearson correlation 0.51, $P = 4.4 \times 10^{-5}$, fig. S16c).

Comparison of technical replicates We compared the alternate allele ratios for heterozygous SNV sites of the 8 inter-array and 4 intra-array replicates. All samples with more than ten sites passing the quality control filtering described above had Pearson correlations that ranged from 0.44 to 0.89, and the Pearson correlation across all replicates was 0.69 ($P = 2.2 \times 10^{-16}$, Fisher and Gayen test, fig. S17). Two of the inter-array samples had only 5 and 7 heterozygous sites with at least 150 reads, respectively, due to low sequencing depth in one of the replicates and were excluded from the comparison. There was otherwise no noticeable difference between inter-array and intra-array replicates. Technical replicates for nonsense variants had moderate correlation (Pearson correlation 0.54, $P = 8.7 \times 10^{-13}$, fig. S18).

Analysis of PTVs The allelic expression of the PTV SNVs was compared to the control sites. The mmPCR-seq data was filtered for quality control purposes using the depth filters described above. The alternate allele ratio was calculated by dividing the number of alternate allele reads by the sum of the reference and alternate allele reads. We observed that high-confidence PTVs have significantly decreased allelic expression of the deleterious allele compared to non-PTV sites ($P < 2.2 \times 10^{-16}$, one-sided t-test).

S7 Insights into nonsense-mediated decay (NMD)

S7.1 Classifying allelic expression patterns

Description of statistical methodology We applied the statistical methods described in Pirinen et al. 2014 (*15*) to all nonsense variants with ASE data (minimum read depth = 8) in the Geuvadis and the GTEx data sets (fig. S19-S26). In GTEx data, this method jointly analyzes ASE data of a given variant across all the tissues per individual. The method takes into account heterogeneity in the data: firstly, individuals often have RNA-sequencing data from varying number of tissues, sometimes from additional tissues than the 9 tissues with the largest sample size; additionally, a given gene is not necessarily expressed in all tissues from which the individual has RNA-seq data; secondly, the total number of RNA-seq reads is variable, with higher uncertainty for sites with low read counts. In a Bayesian modeling framework the uncertainty is propagated throughout the applications of the model to the data. While uncertainty is higher for sites with low read counts, 8 reads carry information and a higher threshold would lead to loss of valuable data (table S4).

Our main motivation are phenomena such as nonsense-mediated decay that are expected to lead to noticeable ASE where one of the alleles may be lowly expressed. Hence, our goal is to classify observed allelic read counts at each site and tissue into one of the following three groups: i) no ASE (group $\mathcal{N}$) where both alleles are (almost) equally expressed, ii) strong ASE (group $\mathcal{S}$) where one of the alleles is expressed very little (if at all), and (iii) moderate ASE (group $\mathcal{M}$) that represents everything in between the first two groups.

In addition, the statistical models allows all variants and all tissues to be studied simultaneously and allows us to address four main questions:

1. Does the variant show patterns of ASE?
2. Which tissues (when multiple tissues are available) show similar ASE effects?
3. What proportion of variants show ASE in all tissues, only in some tissues, or in no tissues?
4. What proportion of variants show very strong ASE effects across all tissues, which are indicative of complete transcript degradation?

We use the following one-sided priors for each of the three groups:

$$\theta(\mathcal{N}) \sim \text{Beta}(2000, 2000)$$
$$\theta(\mathcal{M}) \sim \text{Beta}(36, 12)$$
$$\theta(\mathcal{S}) \sim \text{Beta}(80, 1)$$

Under the no ASE model $\mathcal{N}$ both alleles are expressed (almost) equally and hence $\theta(\mathcal{N}) \approx 0.5$. The $\mathcal{N}$ group model allows small deviations from the exact point value of 0.5 to be robust against technical measurement and mapping bias. Under the moderate ASE model the prior specification dominates at alternate allelic ratios between 0.10 and 0.46. Under the strong ASE model the prior specification dominates at alternate allelic ratio less than 0.1.

In this manuscript, when we apply the model to the GTEx data, we focus on the Hierarchical Grouped Tissue model (GTM*) that allows many variants and tissues to be analyzed simultaneously. On the other hand, when we apply it to single tissue data (Geuvadis), we estimate grouped probabilities with the `gtm` implementation.

In settings where many variants are available for joint analysis the Hierarchical Grouped Tissue model (GTM*) learns from the data the proportion of variants belonging to each of the following five states: 1) NOASE state representing no ASE effects across all tissues analyzed, 2) MODASE state representing moderate ASE effects across all tissues analyzed, 3) SNGASE state representing strong ASE effects across all tissues analyzed, 4) HET0 state representing a mixture of no ASE effects and either moderate and/or strong ASE effects, and 5) HET1 state representing a mixture of moderate and strong ASE effects. The TIS_SPE state shown in the figures represents a sub state of the heterogeneity states. As indicated in Pirinen et al. 2014 (*15*) the default prior specification for the proportions is a dirichlet distribution with hyperparameter vector 1 implying that we are not favoring *a priori* any possible state over the others.

For each individual+tissue pair we estimate the posterior probability that the variant belongs to the no ASE group ($\mathcal{N}$), moderate ASE group ($\mathcal{M}$), and the strong ASE group ($\mathcal{S}$). When we analyze the variants simultaneously, for each variant, we compute the posterior probability that the variant belongs to one of the five states: (N=NOASE, M=MODASE, S=SNGASE, H0=HET0 and H1=HET1), where HET0 is the heterogeneous state with at least one tissue showing no ASE, and HET1 is the heterogeneous state with all tissues showing some ASE (some moderate, some strong). In addition, using the Hierarchical model (GTM*) we obtain estimates of the proportion of variants in each of the five states. We run the ASE models with `nburn=30` and `niter=100`. We report 95% confidence intervals obtained by calculating the 2.5 percentile and 97.5 percentile of the estimated proportions after the 100 iterations.

The method is implemented in the software MAMBA, which is available for download at `http://well.ox.ac.uk/~rivas/mamba/`.

Analysis of RNA-seq data To analyze NMD, we analyzed allelic counts of the following categories of variants:

- SNVs:
    1. synonymous variants (silent);
    2. nonsense variants predicted to escape NMD;
    3. nonsense variants predicted to trigger NMD,
- indels:
    1. in-frame indels;
    2. frameshift indel variants predicted to escape NMD;
    3. frameshift indel variants predicted to trigger NMD.

In the analysis of GTEx data to estimate sharing of ASE across tissues, we analyzed rare variants with minimum number of tissues equal to 2. We used options `two.sided=FALSE` (given our interest in transcript degradation due to premature stop codon) and `indep=FALSE`.

In Figure 2E we claim that the heterogeneous ASE effects observed for the nonsense variant rs149244943 in gene *PHKB* is not driven by a common tissue-specific eQTL. To perform this analysis we checked whether the gene *PHKB* had a single-tissue eQTL as defined in the GTEx main manuscript, and, if so, verified whether the individual in question was homozygous for the top eQTL variant, since cis-regulatory variants can drive ASE only in heterozygous individuals. For this particular example we did not observe any eQTLs for this gene in the GTEx data set.

S7.2 Allelic expression comparison: common versus rare variants

To compare the allelic expression patterns for common (MAF$\geq$ .05) and rare (MAF $\leq$ .01) variants we used estimates of the proportion of variants reported using the ASE module (fig. S19).

We combined the estimates from the multi-tissue model in GTEx as MODASE (moderate effects across all tissues), SNGASE (strong ASE across all tissues), HET0 (mixture of NOASE and/or MODASE, SNGASE) and HET1 (mixture of MODASE and SNGASE) to compare the two categories of variants (common and rare). 95% confidence intervals were obtained from the hierarchical model applied to the data. For Geuvadis, a 95% confidence interval was obtained for the proportion estimates using the normal approximation interval.

We computed a two-sided *P* value for a two-proportion z-test pooled for $H_0 : p_1 = p_2$ where p_1 is proportion of common variants showing no ASE and p_2 is the corresponding proportion of common variants using the `prop.test` function with default parameters (*45-47*).

S7.3 Modeling NMD

The rule for termination-codon position proposed by Nagy and Maquat is: *only those termination codons located more than 50-55 nucleotides upstream of the 3'-most exon-exon junction (measured after splicing) mediates a reduction in mRNA abundance* (*16*).

We used the GTEx ASE outcome as a training data set with binary ASE classification of no ASE (escape; posterior probability $>$ .8) or some form of ASE (trigger; MODASE, SNGASE, HET0, HET1, with sum of the posterior probability $>$ .8) for all nonsense SNVs. We partitioned the data set into a training and a test set using 80% of the data to train the model and 20% to test.

We used the GEUVADIS ASE outcome as an independent test data set with binary ASE classification of no ASE (escape; posterior probability $>$ .8) or moderate/strong ASE (trigger; with sum of the posterior probabilitiy $>$ .8).

We generated a list of 38 sequence and genomic features (table S5) including some used in the development of the CADD approach described in Kircher et al. 2014 (*48*). We fit a model with the 38 predictors. We applied a random forest algorithm using the `caret` package (*49*). To predict the outcome of the independent test data set (Geuvadis) we used the `predict.train` function (from the `caret` package) using the option `type = "prob"` to compute class probabilities. ROC curve was generated using the `pROC` package (*50*) (fig. S21). Importance of features was calculated using the `randomForest` package, which calculated the mean decrease in accuracy and mean decrease in Gini. These statistics were used to rank the 38 features (fig. S22). One of the top ranked features was the distance to the donor site supporting the hypothesis that pre-mRNA splicing is linked to NMD in humans (*51*). Furthermore, the number of downstream exons was ranked above the 50bp rule indicating that the absolute number is an important factor.

S8 Dosage compensation for heterozygous PTVs

Large structural deletions that partially or completely remove genes are confidently expected to cause complete loss of function of the affected genes. Thus, such deletions in addition to nonsense PTVs with strong ASE provide an opportunity to examine the possibility that heterozygous carriers of loss-of-function variants might exhibit compensatory up-regulation of the functional allele, which could contribute to tolerance of PTVs and partially explain the widespread haplosufficiency of human genes (*18*, *52*). In model organisms there are conflicting lines of evidence: some model organism studies appear to indicate clear support for dosage compensation (*19*, *20*) while others appear to indicate that dosage compensation is likely to be unusual (*53*).

A key challenge in the detection of compensation is genotyping error, which is known to be enriched in deletion calls from sequencing data and is expected to produce a signal identical to dosage

compensation because the "heterozygous" individual actually has two functional copies. To minimize the impact of genotyping error on our analyses we focus only on biallelic whole-gene deletion variants with strong experimental support and manual curation.

S8.1 Common large deletions

We first analyzed the few examples of common whole-gene deletion polymorphisms, some of which have been examined in LCLs (*54*). We obtained reliable genotypes for a common deletion of the gene *UGT2B17* in the GTEx donors, and deletions of the genes *DDT*, *GSTT2*, *FAM106A*, *LGALS9C* and *OR2T10* in the GEUVADIS samples (table S6).

S8.2 Rare large deletions

We examined the evidence for signal of dosage compensation in rare deletions in the GTEx and Geuvadis data, analyzing whether the expression levels of heterozygous deletion carriers tend to be half those of the population average. While the raw data show a strong signal of dosage compensation, this signal is largely ablated by LOF annotation filtering and stringent manual curation of CNV genotypes, suggesting a very strong impact of genotyping and annotation error (figs. S27 and S28).

S8.3 Impact of somatic variation

In the DNA data of one GTEx individual we identified a large (20Mb) mosaic deletion. However, careful analysis of the multi-tissue RNA-seq data revealed that the deletion was found only in the individuals blood (where DNA was extracted from), and in other tissues the normal expression of genes spanned by this variant was apparently not due to compensation but by the cells not carrying the somatic deletion (fig. S29).

S8.4 Rare nonsense PTVs with strong ASE

We analyzed whether the gene expression value of rare nonsense PTV carriers with strong ASE signal showed evidence of gene dosage compensation. In the Geuvadis data set we examined a total of 116 nonsense PTVs ($n = 35$ after requiring at least one alternate read observed, fig. S30). In the GTEx data set we examined a total of 25 nonsense PTVs ($n = 18$ after requiring at least one alternate read observed, fig. S31). A total of 53 nonsense PTVs with strong evidence of no genotyping error were used in the final analysis presented in the manuscript.

S9 Insights into impact of variants proximal to splice junction

S9.1 Rare variant analysis

S9.1.1 Statistical Methods: Splice Disruption Model (SDM)

To estimate the impact of rare variants proximal to splice junctions we developed a statistical method we refer to as the **S**plice **D**isruption **M**odel (SDM). In this manuscript we focus on variants in a 50bp window of the donor and the acceptor sites of protein-coding transcripts in the Gencode transcript models. We are interested in the shift of the distribution of splice junction quantification value for carriers of the alternate allele for genetic variants proximal to a splice-junction compared to non-carriers, as a function of distance from splice junction. Using the population values of reads spanning annotated splice junctions and the standardized trait value of the PTV carriers, we estimate, at each distance, the proportion of carriers belonging to

1. the no splice disruption group (0) described by the standard normal distribution, or

2. the splice disruption group (1) with alternative shift in mean μ.

Details: Let us consider all individuals who carry a PTV at a fixed base-pair distance from an acceptor (donor) site of any protein-coding gene. (We do separate analyses for each distance between -25bp and 25bp.) Let y_k be the standardized splice junction quantification value of the PTV carrier k with respect to the population values at the given distance from the acceptor (donor) site. Our SDM is the following mixture model:

$$\begin{aligned} \pi &\sim \text{Beta}(1,1) \\ \gamma_k|\pi &\sim \text{Ber}(\pi) \\ \mu_0 &= 0 \\ \mu_1 &\sim \mathcal{N}(-1,1) \\ \sigma_0^2 &= 1 \\ \sigma_1^2 &\sim \text{IG}(1,1) \\ y_k|\gamma_k,\mu_0,\mu_1 &\sim \mathcal{N}\left(y_k;\mu_{\gamma_k},\sigma_{\gamma_k}^2\right). \end{aligned} \tag{Eq. 2}$$

Motivation of parameters and distribution: The group membership of each of the PTV carriers is unknown in advance. As a result the proportion π of the PTVs belonging to the splice disruption group (characterized by an unknown shift in mean μ_1 and variance σ_1^2) needs to be estimated like the bias of a coin that is estimated from repeated coin tosses. Our prior for π is a uniform distribution on the interval $(0,1)$ (also known as Beta$(1,1)$ distribution) implying that we are not favoring *a priori* any possible value of π over the others.

The trait values have been standardized jointly and we model the trait values of the PTV carriers as either drawn from the general population distribution, i.e. the standard normal distribution $\mathcal{N}(\mu_0=0,\sigma_0^2=1)$, or from a PTV specific normal distribution with unknown shift in mean μ_1 and unknown variance σ_1^2. We model the mean μ_1 by a normal distribution with mean -1 and variance 1 to reflect our interest in those variants that decrease splicing efficiency. In principle, we could also have used another component to reflect variants with a putative increase in splicing efficiency. However, that is beyond the scope of this study. The prior for the variance parameter σ_1^2 is the inverse gamma distribution with parameters $\alpha=1$ and $\beta=1$. This distribution is relatively flat between 0.3 and 1, and thus covers well the region where we expect the variance parameter to reside because the observations have been standardized and altogether have variance of 1.

MCMC algorithm: We use a Gibbs sampler, an approximation algorithm, to analyze the SDM with superscripts for the variables denoting their value after the corresponding iteration.

1. Initialize $\pi^{(0)}, \mu_1^{(0)}, (\sigma_1^2)^{(0)}$, and $\gamma_k^{(0)}$ for all k.

2. Repeat for $t = 1, 2, \ldots, n_{\text{burn}} + n_{\text{iter}}$

   (a) For $k = 1, 2, \ldots, n_{\text{PTV}}$, generate $\gamma_k^{(t)} \sim \text{Ber}\left(p_k^{(t)}\right)$ where

$$p_k^{(t)} = \frac{\pi^{(t-1)}\mathcal{N}\left(y_k;\mu_1^{(t-1)},(\sigma_1^2)^{(t-1)}\right)}{(1-\pi^{(t-1)})\mathcal{N}(0,1)+\pi^{(t-1)}\mathcal{N}\left(y_k;\mu_1^{(t-1)},(\sigma_1^2)^{(t-1)}\right)}. \tag{Eq. 3}$$

(b) Generate $\pi^{(t)} \sim \text{Beta}\left(1+\sum_k \gamma_k^{(t)}, 1+n_{\text{PTV}}-\sum_k \gamma_k^{(t)}\right)$.

(c) Update:

$$(\sigma_1^2)^{(t)} \sim \text{IG}\left(1+\frac{\sum_k \gamma_k^{(t)}}{2}, 1+\frac{1}{2}\sum_k \gamma_k^{(t)}\left(y_k-\mu_1^{(t-1)}\right)^2\right)$$

$$\mu_1^{(t)} \sim \mathcal{N}\left(\frac{\frac{-1}{1}+\frac{\sum_k y_k \gamma_k^{(t)}}{\left(\sigma_1^2\right)^{(t)}}}{\frac{1}{1}+\frac{\sum_k \gamma_k^{(t)}}{\left(\sigma_1^2\right)^{(t)}}}, \left(\frac{1}{1}+\frac{\sum_k \gamma_k^{(t)}}{\left(\sigma_1^2\right)^{(t)}}\right)^{-1}\right).$$

We run the algorithm for $n_{\text{burn}}+n_{\text{iter}}=2000$ iterations and discard the first $n_{\text{burn}}=1000$ iterations from the analysis. We show a typical example of a plot used to evaluate the performance of the algorithm across all the iterations in fig. S32, which were generated for all distances with significant alternatives. We applied the algorithm to the Geuvadis (Fig. 3) and the nine main tissues in the GTEx data set (fig. S33-S37).

Evaluating *P* value: To assess the evidence for the existence of a second component we compute a statistic

$$T^* = \frac{1}{n_{\text{iter}}}\sum_{t=n_{\text{burn}}+1}^{n_{\text{burn}}+n_{\text{iter}}} \pi^{(t)}|\mu_1^{(t)}|. \qquad \text{(Eq. 4)}$$

A standardized statistic is computed

$$T' = T^* \Bigg/ \sqrt{\frac{1}{n_{\text{iter}}-1}\sum_{t=n_{\text{burn}}+1}^{n_{\text{burn}}+n_{\text{iter}}}\left(\pi^{(t)}|\mu_1^{(t)}|-T^*\right)^2} \qquad \text{(Eq. 5)}$$

to calculate an empirical *P* value.

We can calculate an empirical *P* value by calculating null distribution $(T_\ell)_{\ell=1,\ldots,m}$ where for each permutation ℓ, we sample for each splice junction s, where an individual with a PTV is identified, one individual without a PTV, and apply SDM model to the standardized trait values of these sampled individuals. Then, an empirical *P* value is given by

$$p = \frac{\sum_{\ell=1}^{m} I\left(T_\ell \geq T'\right)+1}{m+1}. \qquad \text{(Eq. 6)}$$

S9.1.2 Consequences of splice disruption

In addition to the difficulty in predicting whether splicing changes occur in general, also the type of change - such as exon skipping or elongation - is usually unknown from genetic data alone. Splicing changes always lead to major changes in protein structure either via in-frame changes in exon structure or by introducing a premature stop codon. Our data set provides a valuable opportunity to assess the downstream consequences of variants that disrupt splicing.

To quantify consequences of splice disrupting variants (posterior probability > 0.5 supporting the alternative distribution for sites with $P < .05$ from SDM explained above), we classify the variants into four classes based on the expression values of their proximal introns and exons in the carriers of these variants. As with SDM above, we conduct this analysis separately for each distance within 50bp window around acceptor and donor sites.

Let $y_k = (ex_k, in_k)$ be the pair of standardized expression levels of the proximal exon and intron of the variant carrier k with respect to the population distribution of these quantities. We apply a

Gaussian mixture model that classifies the bivariate expression level values into four groups whose means have prior distributions

$$\begin{aligned}
&\mu_0 = (0,0)^T \qquad \text{(unknown/null)}\\
&\mu_1 \sim \mathcal{N}(m_1 = (-0.5,0)^T, I_2) \qquad \text{(exon skipping)}\\
&\mu_2 \sim \mathcal{N}(m_2 = (0,0.5)^T, I_2) \qquad \text{(exon elongation)}\\
&\mu_3 \sim \mathcal{N}(m_3 = (-0.5,0.5)^T, I_2) \qquad \text{(low exon, high intron; mixture),}
\end{aligned}$$

where I_2 is the two-dimensional identity matrix.
We fix the variance structure of each cluster by setting $\Sigma_0 = I_2$ and $\Sigma_1 = \Sigma_2 = \Sigma_3 = 0.5\, I_2$.
With this notation, the model is

$$\begin{aligned}
\pi &\sim \text{Dirichlet}(1,1,1,1)\\
\gamma_k|\pi &\sim \text{Multinomial}(1,\pi)\\
y_k|\gamma_k,\mu,\Sigma &\sim \mathcal{N}\left(y_k;\mu_{\gamma_k},\Sigma_{\gamma_k}\right). \qquad \text{(Eq. 7)}
\end{aligned}$$

Motivation of parameters and distribution: The group membership of each of the PTV carriers is unknown in advance. As a result the membership between the four groups (group 0: unknown/null, group 1: exon skipping, group 2: exon elongation, and group 3: mixture of exon skipping and exon elongation) needs to be estimated. The likelihood function for the group membership is multinomial. The prior for the proportion vector π of each group is a Dirichlet distribution with equal parameters $\alpha = (1,1,1,1)$ implying that we are not favoring *a priori* any possible value of π over the others.

The bivariate alternative shifts in mean μ_1 (for exon quantification values) and μ_2 (for intron quantification values) are unknown in advance for the PTV carriers. The trait values have been standardized and we model the trait values of the PTV carriers as either drawn from the general population distribution, i.e. $\mathcal{N}\left(m_0 = (0,0)^T, I_2\right)$ or from a PTV specific normal distribution with unknown shift in mean (μ_1, μ_2, or μ_3) and unknown variance matrix (Σ_1, Σ_2, or Σ_3).

We model the exon quantification values of exong skipping group 1 by a normal distribution centered at -0.5 to reflect variants that after splice disruption skip an exon. Analogously, we model the intron quantification values of exong elongation group 2 by a normal distribution centered at 0.5 to reflect variants that after splice disruption elongate an exon. Because these measurements probably do not fully capture the downstream consequences of splice disruption (for instance, some transcripts may have been degraded by NMD) we chose prior distributions that concentrate more mass closer to 0 than in the SDM above to reflect that we believe these values would not be as extreme as the splice junction quantification values that may capture more direct consequences of splice disruption. As indicated in the description of the mixture model, we have fixed the variance structure for each cluster so that we can impose a clear separation between the trait values of the PTV carriers belonging in one of the three alternative groups and the null group.

MCMC algorithm: We use a Gibbs sampler to analyse the model and, in particular, to estimate the proportions π shown in the main manuscript.

Superscripts for the variables denote their value after the corresponding iteration.

1. Initialize $\pi^{(0)}, \mu^{(0)}$, and $\gamma_k^{(0)}$ for all k.

2. Repeat for $t = 1, 2, \ldots, n_{\text{burn}} + n_{\text{iter}}$

   (a) For $k = 1, 2, \ldots, n_{\text{PTV}}$, generate $\gamma_k^{(t)} \sim \text{Multinomial}\left(1, p_k^{(t)}\right)$ where

$$p_{kg}^{(t)} \propto \pi_g^{(t-1)} \mathcal{N}\left(y_k; \mu_g^{(t-1)}, \Sigma_g\right), \text{ for } g = 0,1,2,3.$$

(b) Generate $\pi^{(t)} \sim \text{Dirichlet}\left(1+n_0, 1+n_1, 1+n_2, 1+n_3\right)$, where $n_g = \sum_k I\left(\gamma_k^{(t)} = g\right)$, for $g = 0, 1, 2, 3$.

(c) Update for $g = 1, 2, 3,$:

$$\mu_g^{(t)} \sim \mathcal{N}\left(\left(n_g\Sigma_g^{-1} + I_2\right)^{-1}\left(m_g + n_g\Sigma_g^{-1}\overline{y_g}\right), \left(n_g\Sigma_g^{-1} + I_2\right)^{-1}\right), \text{ where}$$

$$\overline{y_g} = \frac{1}{n_g}\sum_{k:\gamma_k^{(t)}=g} y_k. \qquad \text{(Eq. 8)}$$

We run the algorithm for $n_{\text{burn}} + n_{\text{iter}} = 1010$ iterations and discard the first $n_{\text{burn}} = 10$ iterations from the analysis.

S9.2 Common variant analysis - psiQTLs

To further understand the extent in which alternative splicing is affected by genetic variation, we searched significant associations between common genetic variants (MAF $\geq 5\%$) and exon inclusion (PSI or percentage spliced in) levels in GTEx data (fig. S38-S40, table S7). We calculated the PSI values for 173,483 internal exons from the Gencode v12 annotation. Details are provided in the main GTEx analysis manuscript (*8*). Only exons where the sum of inclusion and exclusion reads is larger or equal to 10 were considered. For each individual we calculated a multi-tissue PSI value by averaging the exons where data for 3 or more tissues is available. We further selected exons with a minimal variability ($\sigma > 0.005$) and searched for significant associations performing Spearman rank correlations. We limited the search of associations (psiQTLs) for common variants that are +/- 25 bp from splice donor or acceptor sites in the vicinity of the exon as these are expected to larger functional impact with eventual PTV to allow direct comparison to rare variant analysis described above (fig. S40). Associations of common genetic variants ($\sim$ 6.1M SNPs (imputed), MAF $\geq 5\%$) to exon PSI values were called psiQTLs. Encode regulatory regions were downloaded from Ensembl and used to perform functional enrichment analysis.

S10 Online resources

PTV results are browseable using the GTEx portal `www.gtexportal.org`. Initially the search entry will be through the `Search Gene Expression` bar by Gene ID or gene symbol. The Protein Truncating Variant (PTV) data are displayed on the GTEx Gene page. The navigation menu on the left side includes an entry "Protein Truncating Variants".

For example visualizations please see `http://kataviz.github.io/html/ase.html` and `http://kataviz.github.io/html/ptv.html`.

Improved predictive models for NMD are supported in the MAMBA software `http://www.well.ox.ac.uk/~rivas/mamba/` and will be updated with each data release.

S11 Consortia members

The GTEx consortium

Analysis working group: LDACC Kristin G. Ardlie[1], David S. Deluca[1], Ayellet V. Segre[1], Timothy J. Sullivan[1], Taylor R. Young[1], Ellen T. Gelfand[1], Casandra A. Trowbridge[1], Julian B. Maller[1,2], Taru Tukiainen[1,2], Monkol Lek[1,2], Lucas D. Ward[1,3], Pouya Kheradpour[1,3], Benjamin Iriarte[3], Yan Meng[1], Cameron D. Palmer[1,4], Tonu Esko[1,4,5], Wendy Winckler[1], Joel Hirschhorn[1,4], Manolis Kellis[1,3], Daniel G. MacArthur[1,2], Gad Getz[1,6]; **UNC/NCSU** Andrey A. Shablin[7], Gen Li[8], Yi-Hui Zhou[9], Andrew B. Nobel[8], Ivan Rusyn[10,11], Fred A. Wright[9]; **U Geneva** Tuuli Lappalainen[12,13,14,15,16,17], Pedro G. Ferreira[12,13,14], Halit Ongen[12,13,14], Manuel A. Rivas[18], Alexis Battle[19,20], Sara Mostafavi[19], Jean Monlong[21,22,23], Michael Sammeth[21,22,24], Marta Mele[21,22,25], Ferran Reverter[21,26], Jakob Goldmann[21,27], Daphne Koller[19], Roderic Guigo[21,22,28], Mark I. McCarthy[18,29,30], Emmanouil T. Dermitzakis[12,13,14]; **U Chicago** Eric R. Gamazon[31,32], Hae Kyung Im[31], Anuar Konkashbaev[31,32], Dan L. Nicolae[31], Nancy J. Cox[31,32]; Timothe Flutre[33,34], Xiaoquan Wen[35], Matthew Stephens[33,36], Jonathan K. Pritchard[33,37,38]; **Harvard** Zhidong Tu[39,40], Bin Zhang[39,40], Tao Huang[39,40], Quan Long[39,40], Luan Lin[39,40], Jialiang Yang[39,40], Jun Zhu[39,40], Jun Liu[41].

Biospecimen and data collection, processing, quality control, storage, and pathological review: caHUB Biospecimen Source Sites: NDRI Amanda Brown[42], Bernadette Mestichelli[42], Denee Tidwell[42], Edmund Lo[42], Mike Salvatore[42], Saboor Shad[42], Jeffrey A. Thomas[42], John T. Lonsdale[42]; **Roswell Park** Christopher Choi[43], Ellen Karasik[43], Kimberly Ramsey[43], Michael T. Moser[43], Barbara A. Foster[43], Bryan M. Gillard[43]; **Science Care Inc.** John Syron[44], Johnelle Fleming[44], Harold Magazine[44]; **Gift of Life Donor Program** Rick Hasz[45]; **LifeNet Health** Gary D. Walters[46]; **UNYTS** Jason P. Bridge[47], Mark Miklos[47], Susan Sullivan[47]. **caHUB ELSI Study: VCU** Laura K. Barker[48], Heather Traino[48], Magboeba Mosavel[48], Laura A. Siminoff[48,49]. **caHUB Comprehensive Biospecimen Resource: Van Andel Research Institute** Dana R. Valley[50], Daniel C. Rohrer[50], Scott Jewel[50]. **caHUB Pathology Resource Center: NCI** Philip Branton[51]; **Leidos Biomedical Research Inc.** Leslie H. Sobin[52], Mary Barcus[52]. **caHUB Comprehensive Data Resource: Leidos Biomedical Research Inc.** Liqun Qi[52], Pushpa Hariharan[52], Shenpei Wu[52], David Tabor[52], Charles Shive[52].

caHUB Operations Management: Leidos Biomedical Research Inc. Anna M. Smith[52], Stephen A. Buia[52], Anita H. Undale[52], Karna L. Robinson[52], Nancy Roche[52], Kimberly M. Valentino[52], Angela Britton[52], Robin Burges[52], Debra Bradbury[52], Kenneth W. Hambright[52], John Seleski[53], Greg E. Korzeniewski[52]; **Sapient Government Services** Kenyon Erickson[54].

Brain Bank Operations: University of Miami Yvonne Marcus[55], Jorge Tejada[55], Mehran Taherian[55], Chunrong Lu[55], Barnaby E. Robles[55], Margaret Basile[55], Deborah C. Mash[55].

Program Management: NHGRI Simona Volpi[56], Jeffery P. Struewing[56], Gary F. Temple[56], Joy Boyer[57], Deborah Colantuoni[56]; **NIMH** Roger Little[58], Susan Koester[59]; **NCI** Latarsha J. Carithers[51], Helen M. Moore[51], Ping Guan[51], Carolyn Compton[51], Sherilyn J. Sawyer[51], Joanne P. Demchok[60], Jimmie B. Vaught[51], Chana A. Rabiner[51], Nicole C. Lockhart[55,57].

[1]The Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02142, USA. [2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [3]MIT Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [4]Center for Basic and Translational Obesity Research and Division of Endocrinology, Boston Children's Hospital, Boston, Massachusetts 02115, USA. [5]Estonian Genome Center, University of Tartu, Tartu, Estonia. [6]Cancer Center and Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, 02114, USA. [7]Center for Biomarker Research and Personalized Medicine, Virginia Commonwealth University, Richmond, Virginia 23298, USA. [8]Department of Statistics and Operations Research and Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599, USA. [9]Bioinformatics Research Center and Departments of Statistics and Biologi-

cal Sciences, North Carolina State University, Raleigh, North Carolina 27695, USA. [10]Department of Environmental Sciences and Engineering, University of North Carolina, Chapel Hill, NC 27599. [11]Department of Veterinary Integrative Biosciences, Texas AM University, College Station, Texas 77843, USA. [12]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland. [13]Institute for Genetics and Genomics in Geneva (iG3), University of Geneva, 1211 Geneva, Switzerland. [14]Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland. [15]Department of Genetics, Stanford University, Stanford, California 94305, USA. [16]New York Genome Center, New York, New York 10011, USA. [17]Department of Systems Biology, Columbia University Medical Center, New York, New York 10032, USA. [18]Wellcome Trust Centre for Human Genetics Research, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, United Kingdom OX3 7BN. [19]Department of Computer Science, Stanford University, Stanford, California 94305, USA. [20]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland 21218, USA. [21]Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain. [22]Universitat Pompeu Fabra, 08003 Barcelona, Catalonia, Spain. [23]Human Genetics Department, McGill University, H3A 0G1 Montral Canada. [24]National Institute for Scientific Computing, Petropolis 25651-075 Rio de Janeiro, Brazil. [25]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [26]Universitat de Barcelona, 08028 Barcelona, Catalonia, Spain. [27]Radboud University Nijmegen, Netherlands. [28]Institut Hospital del Mar d'Investigacions Mdiques (IMIM), 08003 Barcelona, Spain. [29]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford, United Kingdom OX3 7LJ. [30]Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, United Kingdom OX3 7LJ. [31]Section of Genetic Medicine, Department of Medicine and Department of Human Genetics, University of Chicago, Chicago, Illinois 60637. [32]Division of Genetic Medicine, Department of Medicine, Vanderbilt University, Nashville, TN 37232. [33]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA. [34]INRA, Department of Plant Biology and Breeding, UMR 1334, AGAP, Montpellier, 34060, France. [35]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA. [36]Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA. [37]Department of Genetics and Biology, Stanford University, Stanford, California 94305, USA. [38]Howard Hughes Medical Institute, Chicago, Illinois, USA. [39]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [40]Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [41]Department of Statistics, Harvard University, Cambridge, Massachusetts 02138. [42]National Disease Research Interchange, Philadelphia, Pennsylvania 19103, USA. [43]Roswell Park Cancer Institute, Buffalo, New York 14263, USA. [44]Science Care, Inc., Phoenix, Arizona, USA. [45]Gift of Life Donor Program, Philadelphia, Pennsylvania 19103, USA. [46]LifeNet Health, Virginia Beach, Virginia 23453, USA. [47]UNYTS, Buffalo, New York 14203, USA. [48]Virginia Commonwealth University, Richmond, Virginia 23298, USA. [49]Department of Public Health, Temple University, Philadelphia, Pennsylvania 19122, USA. [50]Van Andel Research Institute, Grand Rapids, Michigan 49503. [51]Biorepositories Biospecimen Research Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [52]Biospecimen Research Group, Clinical Research Directorate, Leidos Biomedical Research, Inc., Rockville, Maryland 20852, USA. [53]iDoxSolutions, Inc., Bethesda, MD 20814 [54]Sapient Government Services, Arlington, Virginia 22201-2909. [55]Brain Endowment Bank, Department of Neurology, Miller School of Medicine, University of Miami, Miami, Florida 33136, USA. [56]Division of Genomic Medicine, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [57]Division of Genomics and Society, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [58]Office of Science Policy, Planning, and Communications, National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland 20892, USA. [59]Division of Neuroscience and Basic Behavioral Science, National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland 20892, USA. [60]Cancer Diagnosis Program,

National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.

Geuvadis consortium

Tuuli Lappalainen[1,2,3], Michael Sammeth[4,5], Marc R Friedlander[5,6], Peter AC 't Hoen[7], Jean Monlong[5], Manuel A Rivas[8], Mar Gonzlez-Porta[9], Natalja Kurbatova[9], Thasso Griebel[4], Pedro G Ferreira[5,6], Matthias Barann[10], Thomas Wieland[11], Liliana Greger[9], Maarten van Iterson[7], Jonas Almlof[12], Paolo Ribeca[4], Irina Pulyakhina[7], Daniela Esser[10], Thomas Giger[1], Andrew Tikhonov[9], Marc Sultan[13], Gabrielle Bertier[5,6], Daniel G MacArthur[14,15], Monkol Lek[14,15], Esther Lizano[5,6], Henk PJ Buermans[7,16], Ismael Padioleau[1,2,3], Thomas Schwarzmayr[11], Olof Karlberg[12], Halit Ongen[1,2,3], Helena Kilpinen[1,2,3], Sergi Beltran[4], Marta Gut[4], Katja Kahlem[4], Vyacheslav Amstislavskiy[13], Oliver Stegle[9], Matti Pirinen[8], Stephen B Montgomery[1], Peter Donnelly[8], Mark I McCarthy[8,17], Paul Flicek[9], Tim M Strom[11,18], Hans Lehrach[13,19], Stefan Schreiber[10], Ralf Sudbrak[13,19], Angel Carracedo[20], Stylianos E Antonarakis[1,2], Robert Hasler[10], Ann-Christine Syvanen[12], Gert-Jan van Ommen[7], Alvis Brazma[9], Thomas Meitinger[11,18], Philip Rosenstiel[10], Roderic Guigo[5,6], Ivo G Gut[4], Xavier Estivill[5,6], Emmanouil T Dermitzakis[1,2,3]

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland

[2]Institute for Genetics and Genomics in Geneva (1G3), University of Geneva, 1211 Geneva, Switzerland

[3]Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland

[4]Centro Nacional d'Anlisi Genmica, 08028 Barcelona, Spain

[5]Center for Genomic Regulation (CRG), 08003 Barcelona, Spain

[6]Pompeu Fabra University (UPF), 08003 Barcelona, Spain

[7]Department for Human and Clinical Genetics, Leiden University Medical Center, 2300 RC Leiden, the Netherlands

[8]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom

[9]European Bioinformatics Institute, EMBL-EBI, Hinxton, United Kingdom

[10]Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, D-24105 Kiel, Germany

[11]Institute of Human Genetics, Helmholtz Zentrum Mnchen, 85764 Neuherberg, Germany

[12]Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, 751 85 Uppsala, Sweden

[13]Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany

[14]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA

[15]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge MA 02142, USA

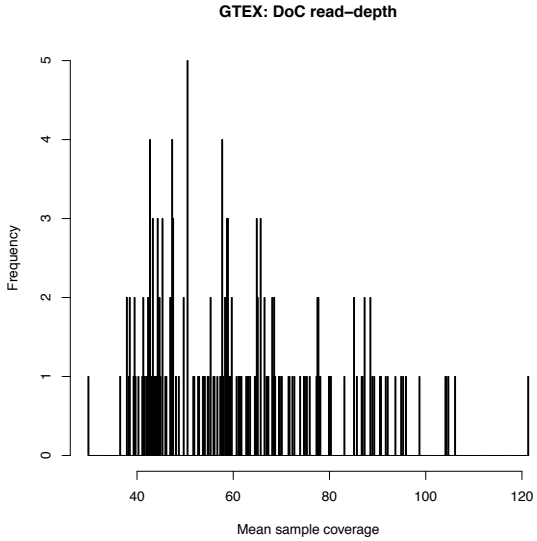
[16]Leiden Genome Technology Center, 2300 RC Leiden, the Netherlands

[17]Oxford Centre for Diabetes Endocrinology and Metabolism, University of Oxford, Oxford OX3 7BN, United Kingdom

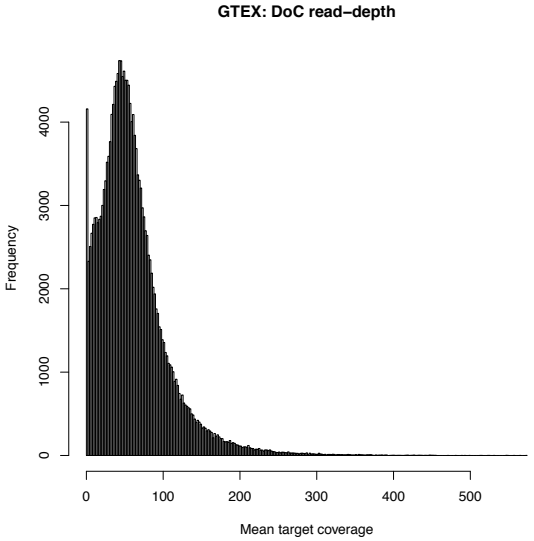
[18]Institute of Human Genetics, Technische Universitt Mnchen, 81675 Munich, Germany

[19]Dahlem Centre for Genome Research and Medical Systems Biology, 14195 Berlin, Germany

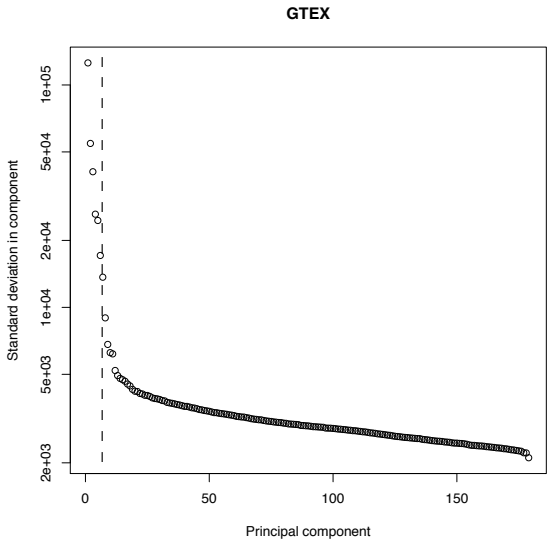
[20]Fundacion Publica Galega de Medicina Xenomica SERGAS, Genomic Medicine Group CIBERER, Universidade de Santiago de Compostela, Santiago de Compostela, Spain



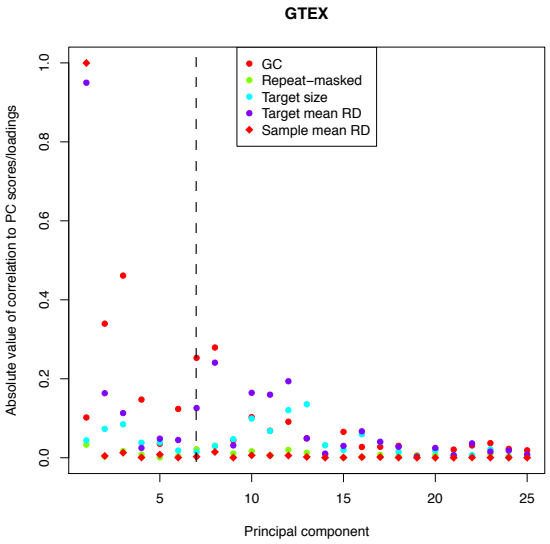
(a) Mean sample coverage



(b) Mean target coverage

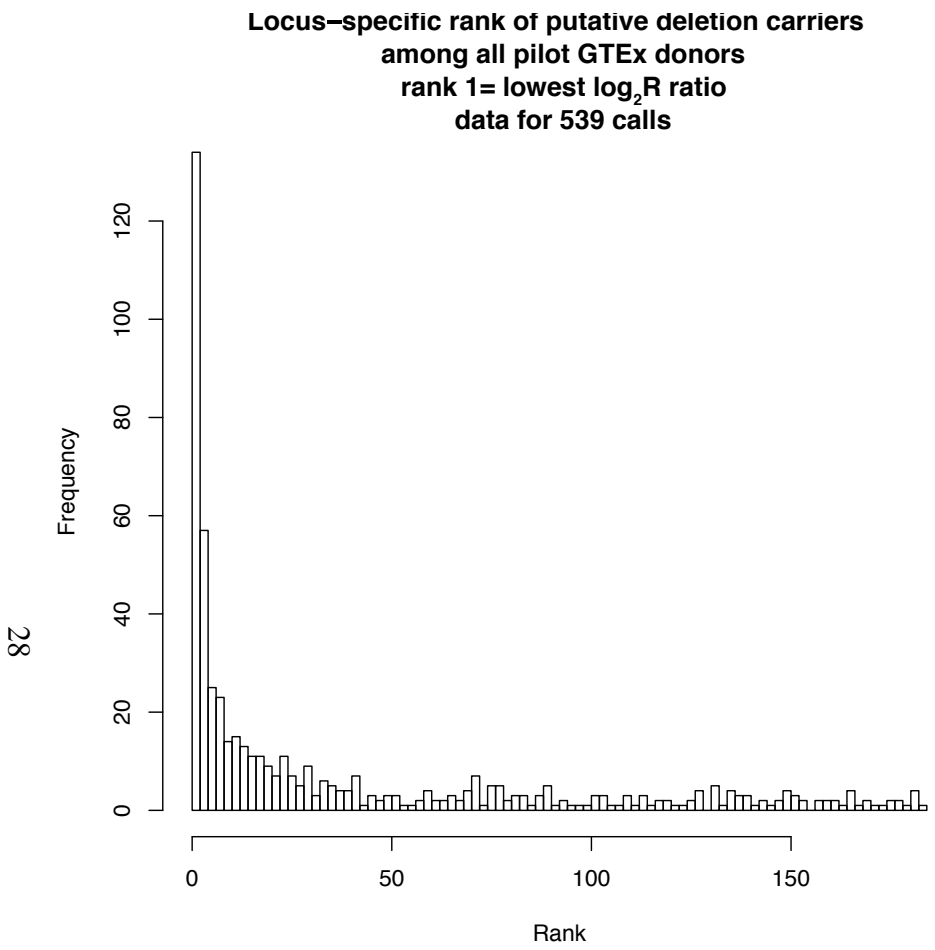


(c) Principal components

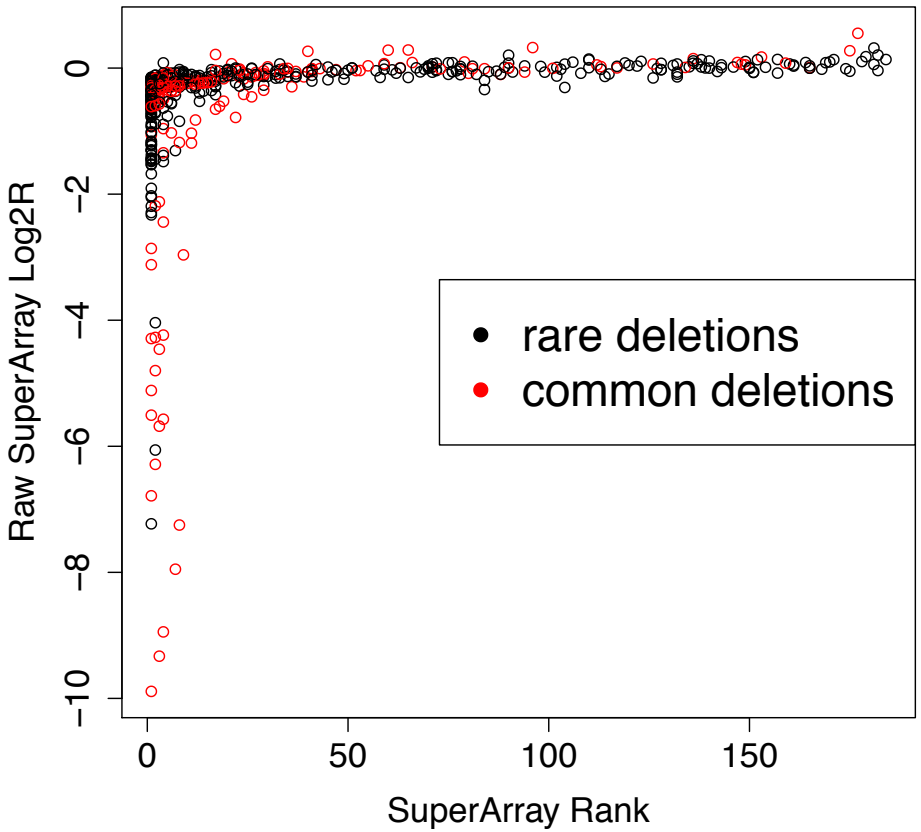


(d) PC correlations

Figure S1: CNV quality control in the GTEx exome sequencing data set using XHMM. a) Coverage was calculated and averaged over each target using GATKDepthOfCoverage. Mean sample coverage across the exome ranged between 40x and 80x. b) Approximately 4000 targets were found to have little or no coverage in most samples. c) Using XHMM, normalization was performed in a principal component analysis framework, where the 7 largest principal components were automatically removed. d) The largest components were correlated with a combination of GC content, and sample and target read depths.



(a) Validation of XHMM calls (1)



(b) Validation of XHMM calls (2)

Figure S2: Validation of XHMM large deletion calls using array data. a) For each of the 399 deletion loci in our exome-based rare CNV map, we summarize the copy number of each individual as the mean $\log_2$ ratio of SuperArray probes in the locus. We then calculated the rank of those $\log_2$ ratios, resulting in 184 ranks for each locus. This histogram shows the distribution of ranks corresponding to all XHMM deletion calls. Individuals with XHMM calls tend to be low ranks at called sites (lower rank = lower $\log_2$ ratio). b) Here we show the relationship between SuperArray rank and $\log_2$ ratio for all XHMM calls. We have colored all calls based on whether or not they occur at known deletion loci previously described by the 1000 Genomes project or Structural Variation Consortium.

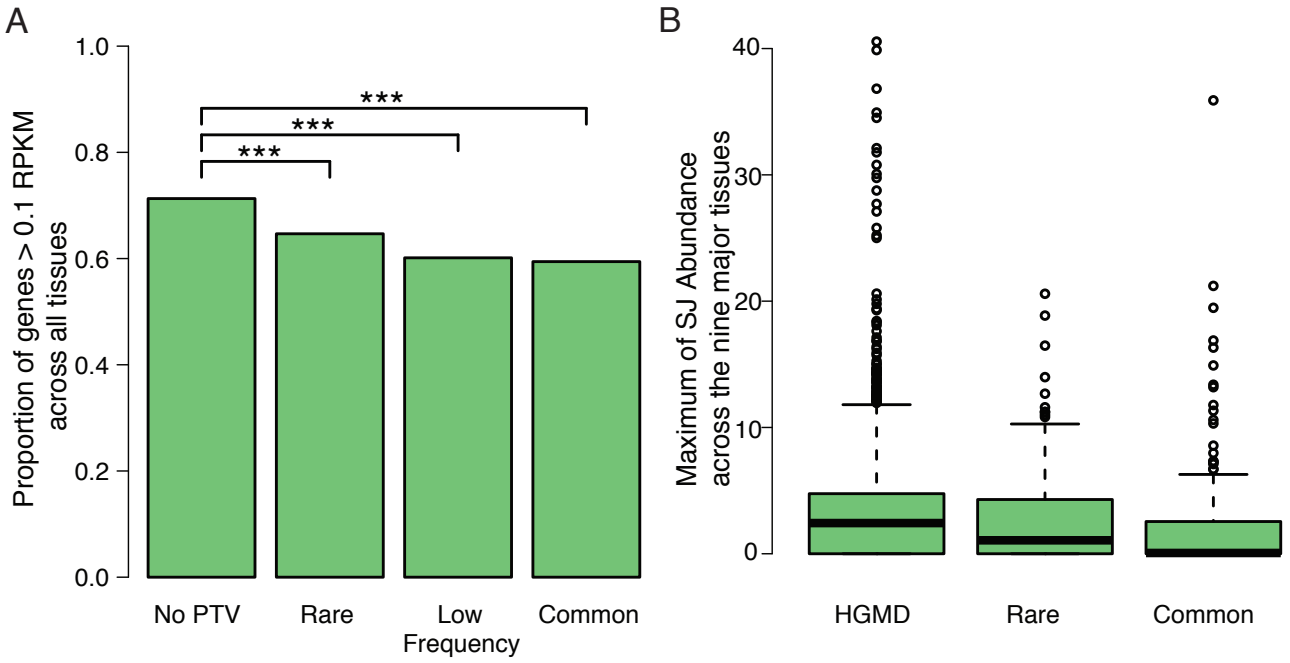


Figure S3: Transcriptional properties of PTV containing transcripts: ubiquitous expression and splice junction usage. a) Proportion of ubiquitously expressed genes, defined as median gene expression level (calculated across individuals) > 0.1 RPKM for all nine tissues in GTEx, in protein-coding genes where no PTVs, and genes with PTVs of different frequency: common (MAF $\geq .05$), low frequency ($.01 <$ MAF $< .05$), and c) rare (MAF $\leq .01$). All three categories of PTV-containing genes were less likely to be ubiquitously expressed compared to the set of protein-coding genes with no PTVs ($P < 1.9 \times 10^{-10}$, two-proportion z-test). b) Maximum of the medians splice junction abundance across the nine tissues in junctions containing splice-disrupting variants that are common, rare, and present in the Human Genome Mutation Database (HGMD) across the nine tissues. The junction abundance is measured from individuals not carrying the variants. Junctions with common splice-disrupting variants are less often used compared to those containing rare splice-disrupting variants (maximum SJ comparison $P = 0.0015$, MWW test) and reported disease-causing variants in HGMD (SJ comparison $P < 3.3 \times 10^{-12}$, MWW test).

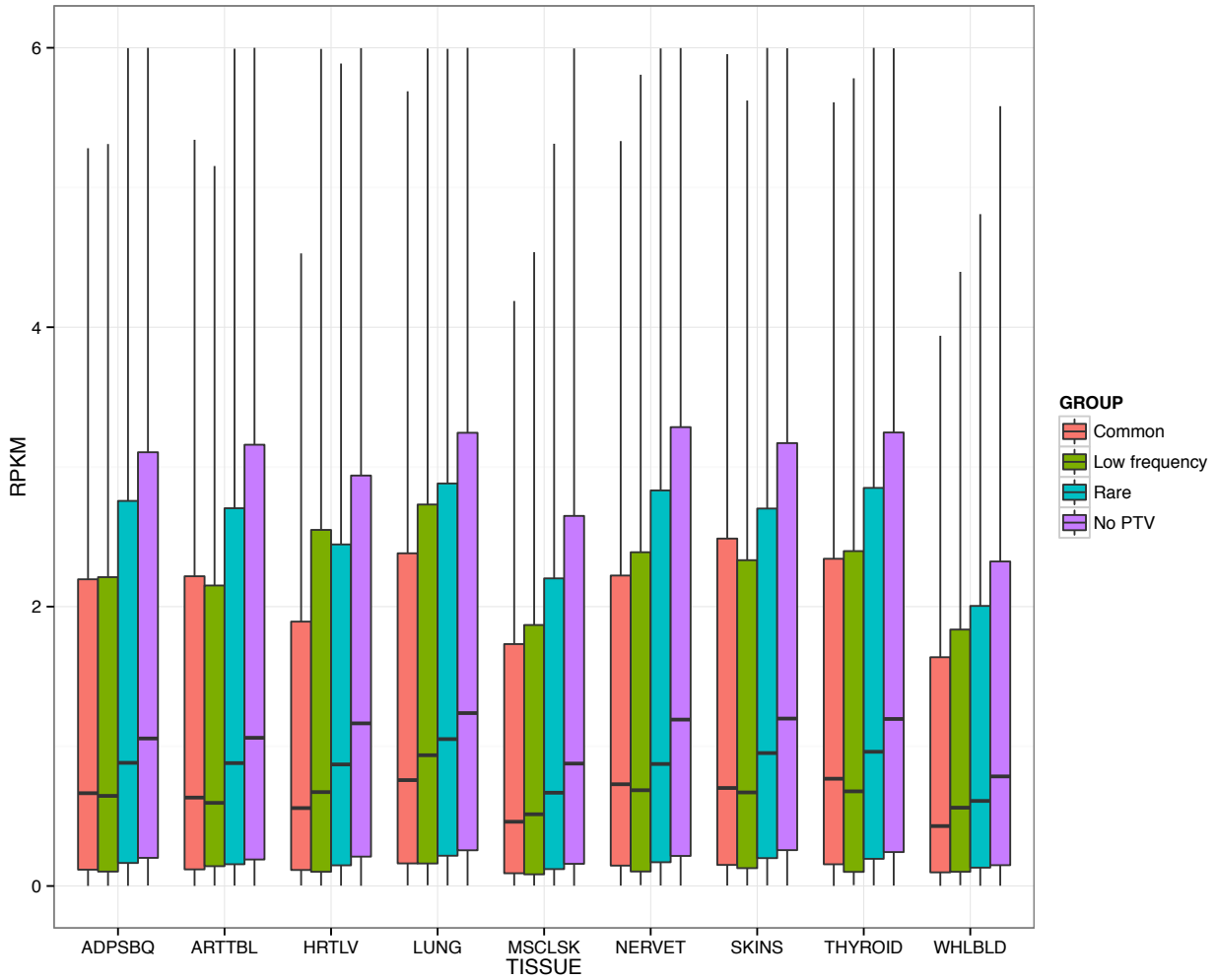
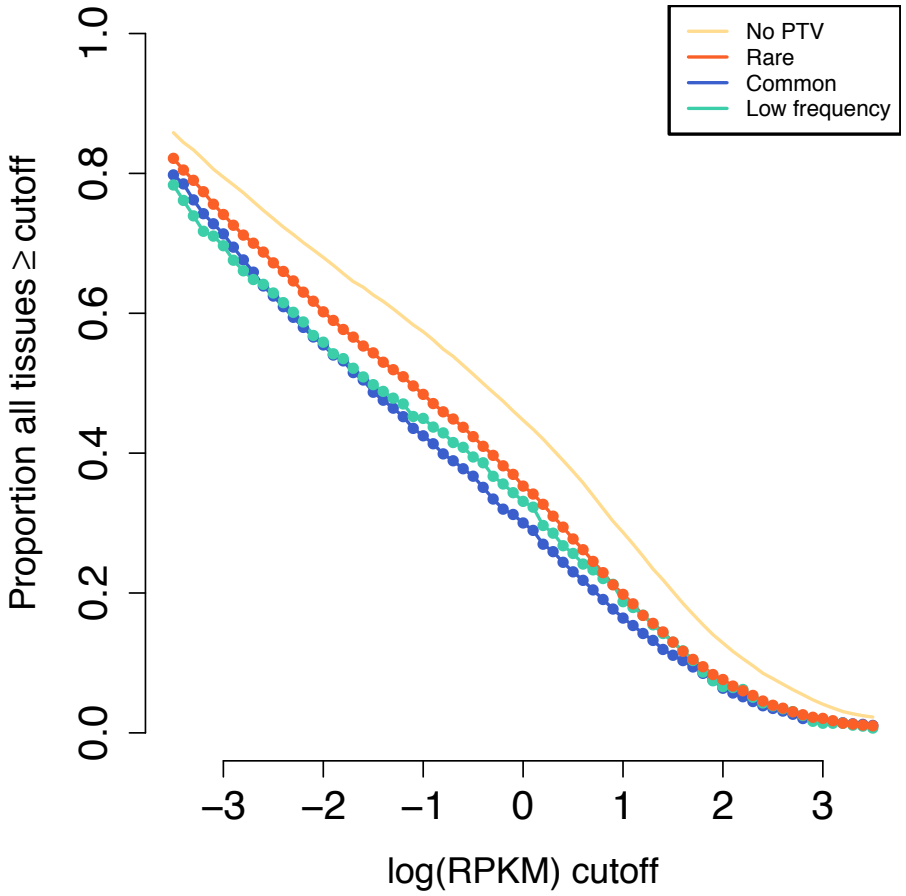
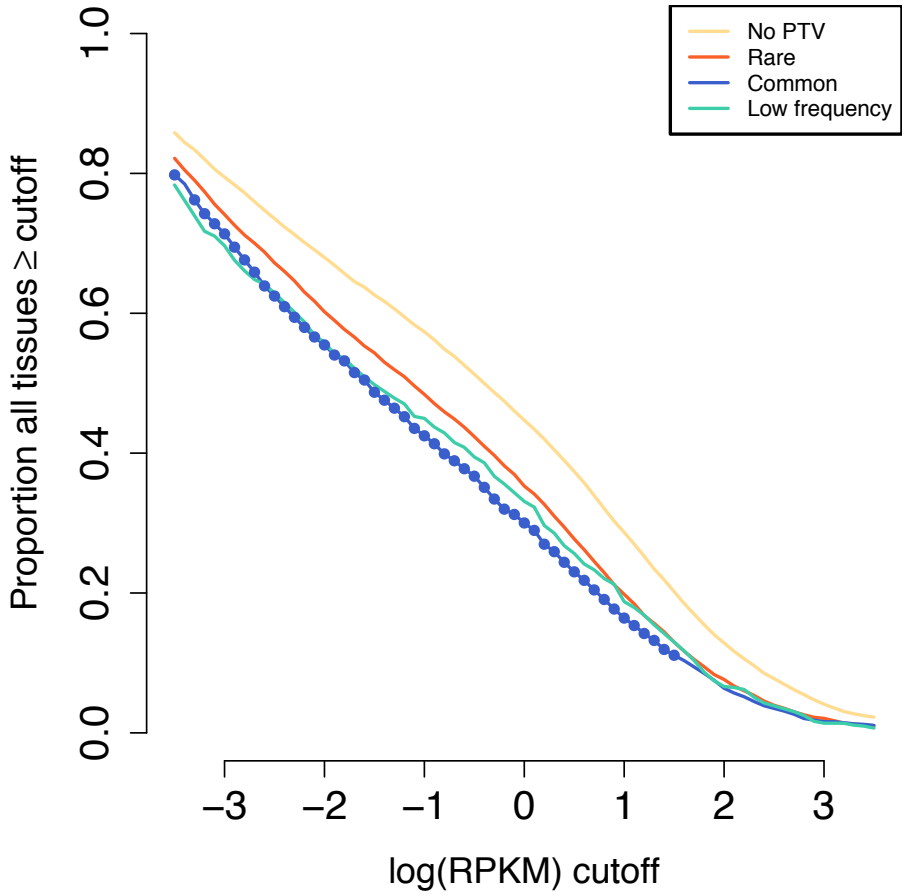


Figure S4: Transcriptional properties of PTV containing transcripts: tissue-wide expression profile for PTV containing genes. Comparison of the distribution of median gene expression values for PTV containing genes across tissues and categories. We find evidence that genes containing common PTVs are consistently lower expressed across tissues compared to genes that do not contain PTVs in these data sets (*P* value per tissue $< 2 \times 10^{-16}$, MWW test) and the same is observed also for genes containing low-frequency or rare PTVs (*P* value per tissue 6.1×10^{-7} to 2.4×10^{-11} and $< 2 \times 10^{-16}$, MWW test).

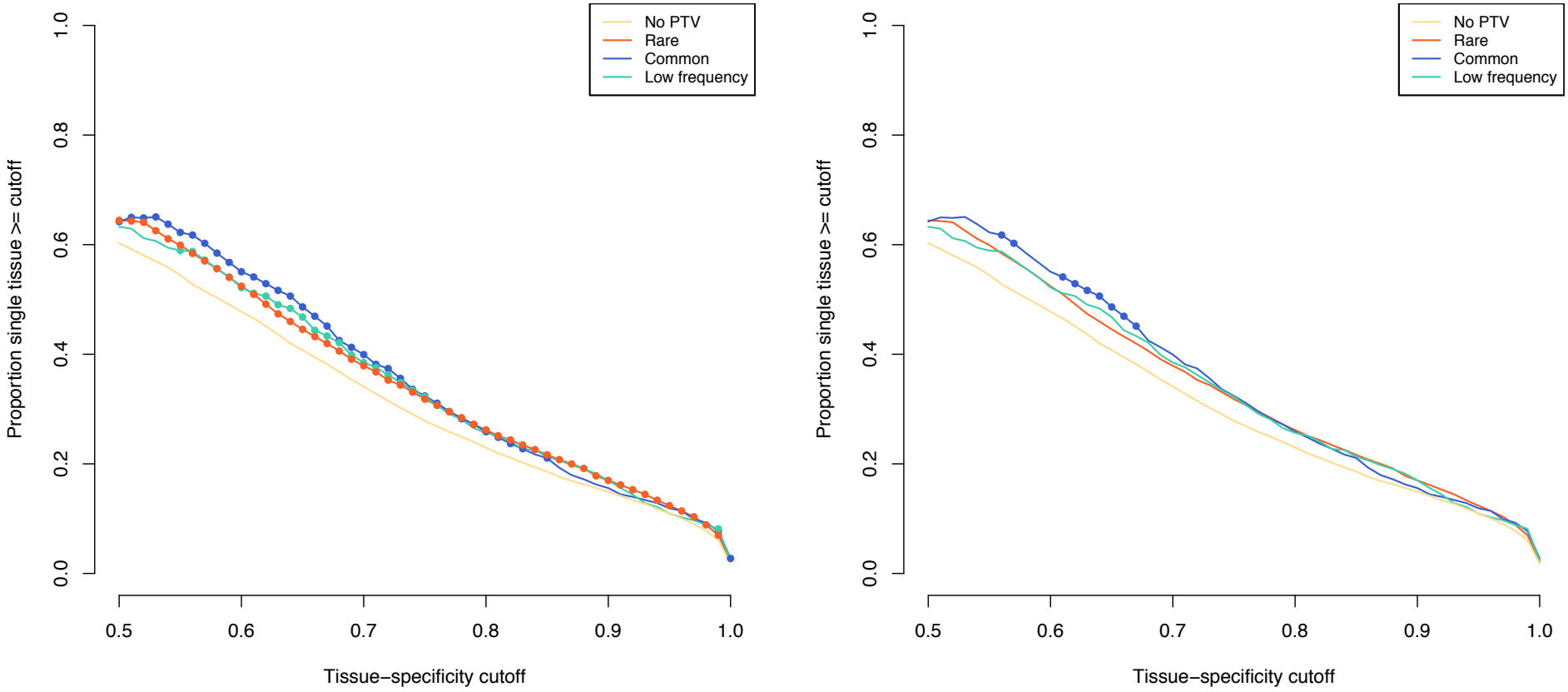


(a) Comparison to all protein-coding genes



(b) Comparison to rare

Figure S5: Transcriptional properties of PTV containing transcripts: assessment of ubiquitous expression for PTV containing genes. Comparison of the proportion of genes with median gene expression value above a log(RPKM) cutoff for rare (orange), common (blue), low frequency (aqua), and no PTV containing protein-coding genes (yellow). We find a) evidence that PTV-containing genes are less likely to be ubiquitously expressed [filled circles denote two-proportion test $P < .05$] compared to the protein-coding genes that do not contain PTVs in these data sets (no PTV) and b) evidence that genes containing common PTVs are less likely to be ubiquitously expressed than rare PTV containing genes.



(a) Comparison to all protein-coding genes

(b) Comparison to rare

Figure S6: Transcriptional properties of PTV containing transcripts: assessment of tissue-specific expression for PTV containing genes. Comparison of the proportion of genes above a tissue-specificity cutoff for rare (orange), common (blue), low frequency (aqua), and no PTV containing protein-coding genes (yellow). We find a) evidence that common PTV containing genes are more likely to be tissue-specific compared to protein-coding genes that do not contain PTVs in these data sets (no PTV) [filled circles denote two-proportion test $P < .05$]. b) Occasionally we observe significant differences in the tissue-specific expression comparison between common and rare PTV containing genes (blue filled circles).

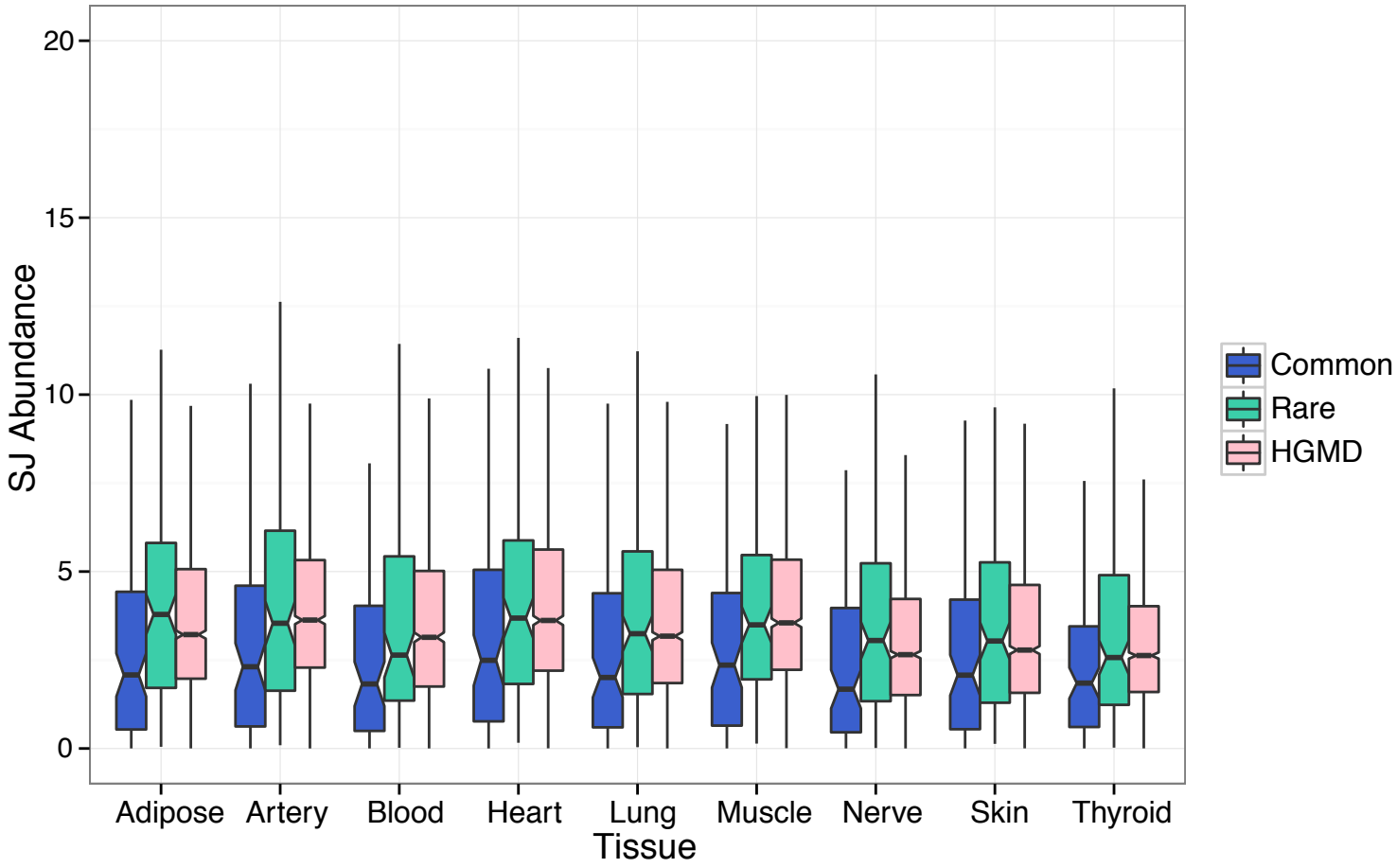


Figure S7: Transcriptional properties of PTV containing transcripts: comparison of splice junction usage for splice-disrupting variant containing junctions. For each of the tissues a boxplot is plotted using the median splice junction abundance value in homozygous reference allele carriers for splice junctions that are impacted by an annotated splice-disrupting PTVs. We observe that junctions with common annotated splice-disrupting variants are less often used compared to those containing rare splice-disrupting variants across all studied tissues ($P = 0.01$ to 1.4×10^{-4}, MWW test). They were also less likely to be used compared to junctions with splice-disrupting variants in HGMD.

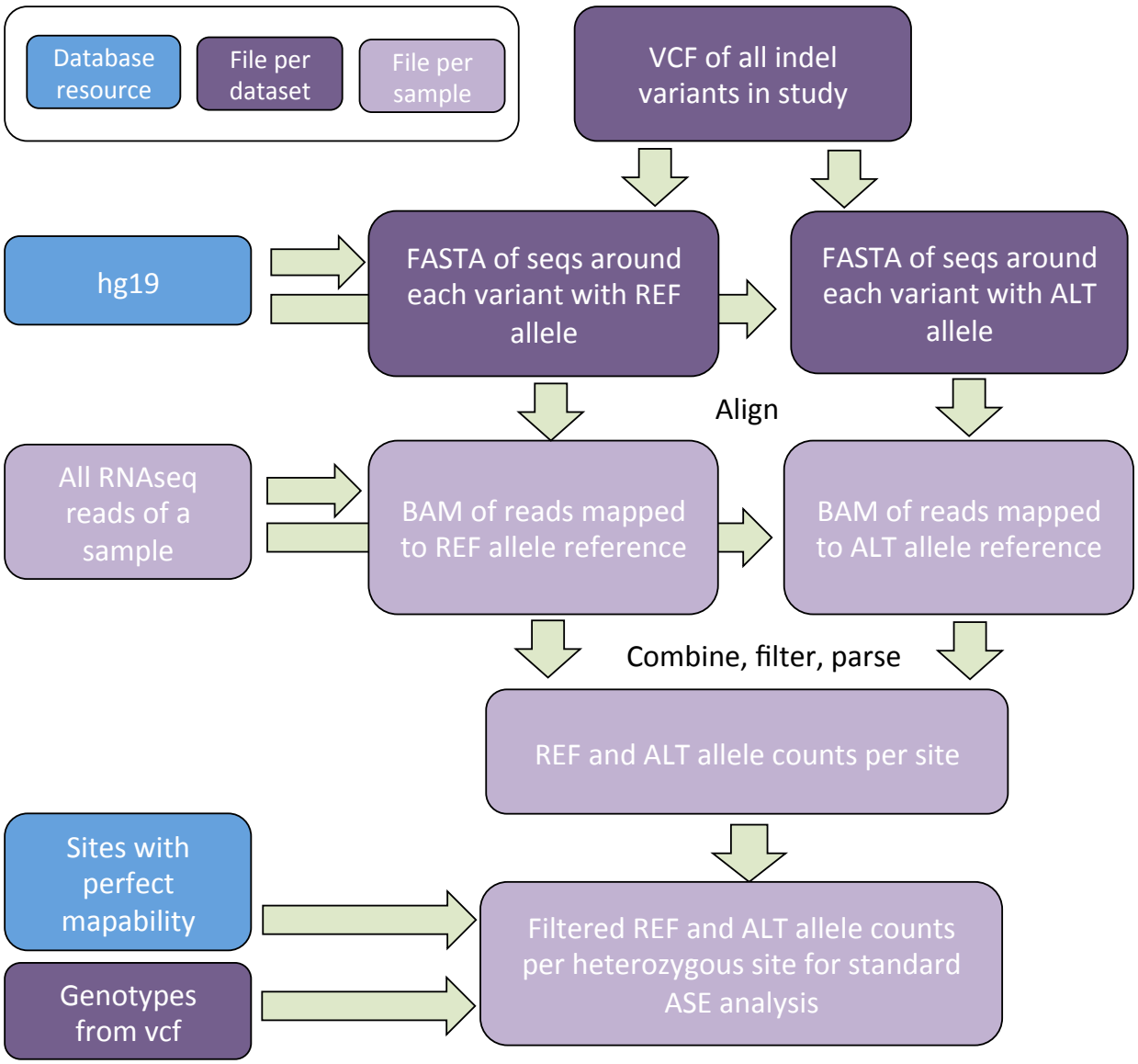


Figure S8: Allele-specific expression: schematic diagram of the indel ASE pipeline. Our method is based on alignment to local reference sequences that have been modified to contain both the reference and alternative alleles. Briefly, for each variant we extract the flanking ($\pm$ 100bp) reference genomic sequence, and modify it to build an alternative allele reference index. The RNA-seq reads are aligned to both of the indexes separately, and final allele counts are retrieved from their combination.

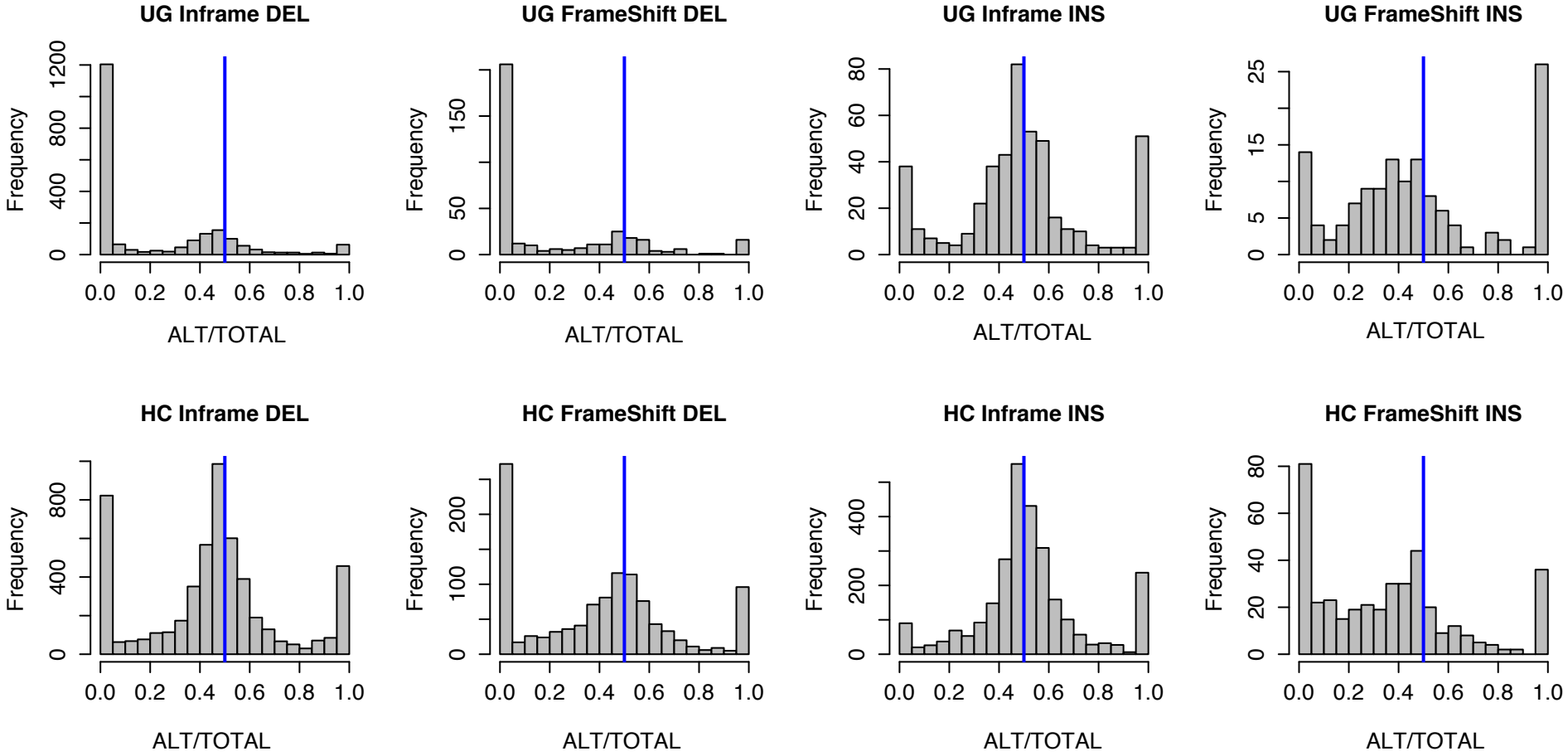


Figure S9: Allele-specific expression: RNA-seq based quality estimation of heterozygous indel genotypes called with GATK UnifiedGenotyper (UG; first row) or GATK HaplotypeCaller (HC; second row). The histograms show the alternate allele ratio observed in RNA-seq data for genotypes called as heterozygous from exome-sequencing data, in different variant classes of in-frame deletions, frameshift deletions, in-frame insertions, and frameshift insertions (columns left to right). The expected allele ratio would be around 0.5 for true heterozygous genotypes, or slightly shifted to the left for frameshift variants due to NMD. Observing the poor quality of especially deletion genotypes from GATK UnifiedGenotyper, we decided to use genotypes from GATK HaplotypeCaller for the indel analysis.

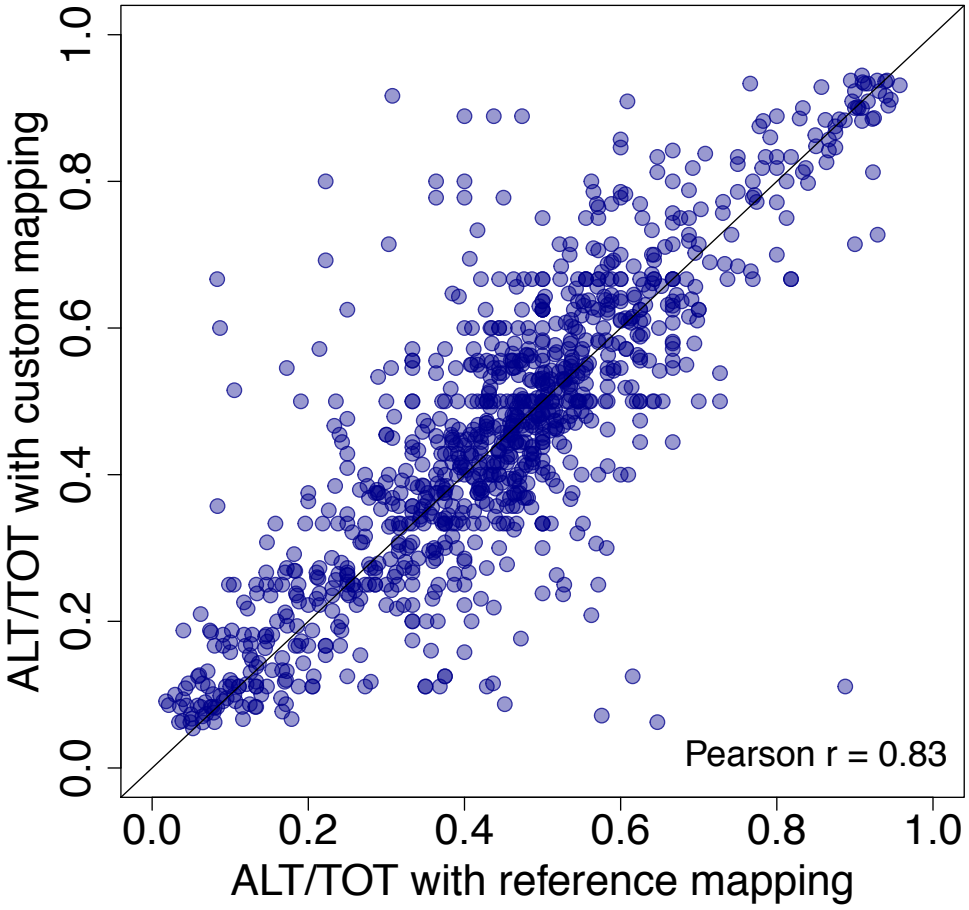
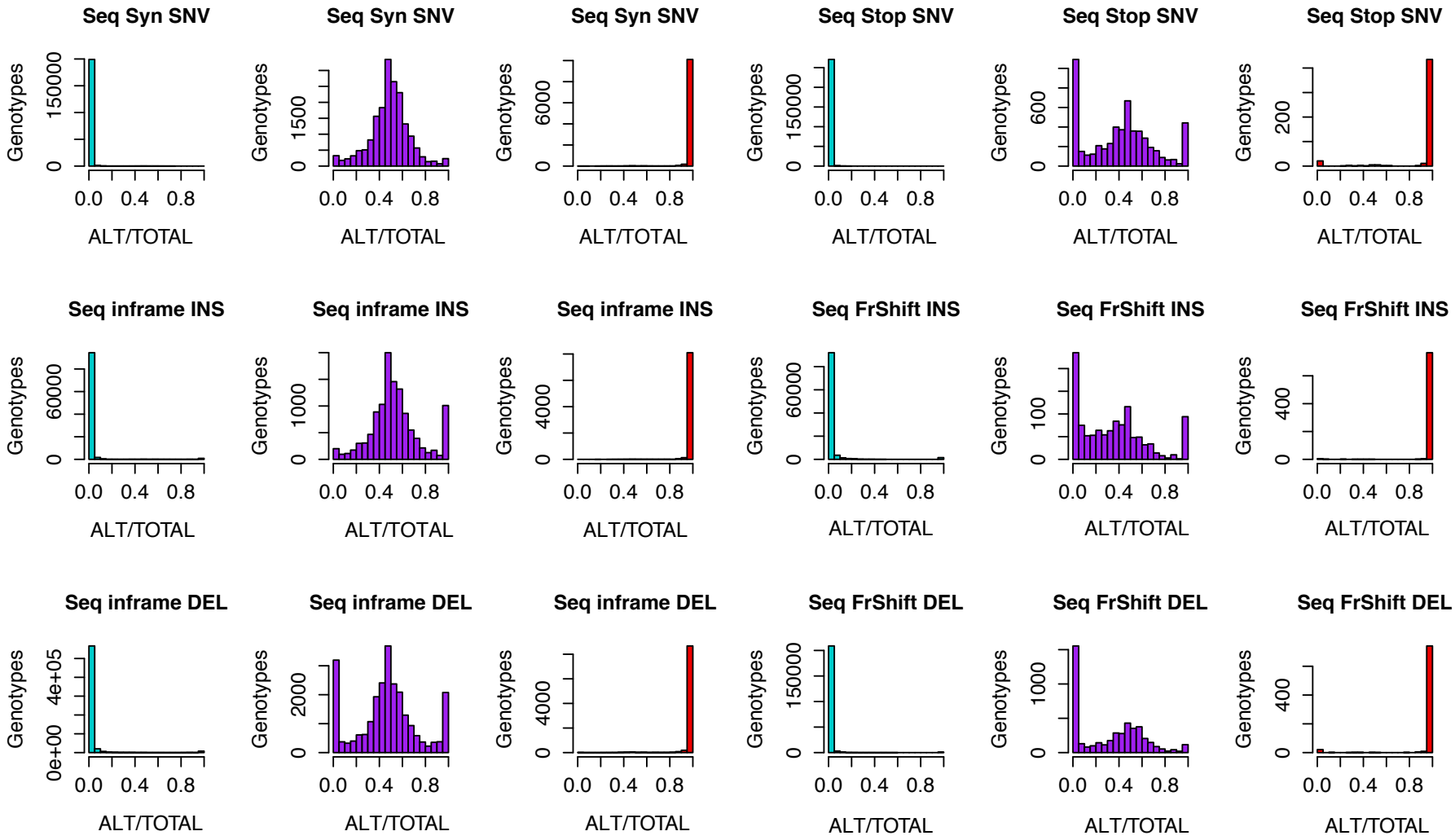


Figure S10: Allele-specific expression: comparison of allelic ratios of SNVs analyzed with the standard ASE pipeline with mapping to the reference genome, and the new ASE pipeline with mapping to local customized references.



(a) Control variants

(b) PTVs

Figure S11: Allele-specific expression: allelic ratios of all genotypes, with REF/REF in cyan, REF/ALT in purple, and ALT/ALT in red, for different variant classes with the PTV classes (nonsense SNVs, frameshift insertions, frameshift deletions) shown on the right and the corresponding controls (synonymous SNVs, in-frame insertions, in-frame deletions) on the left. Note that as the histograms include all measured genotypes, common variants become overrepresented and thus these results capture genotype quality rather than an unbiased signal of NMD.

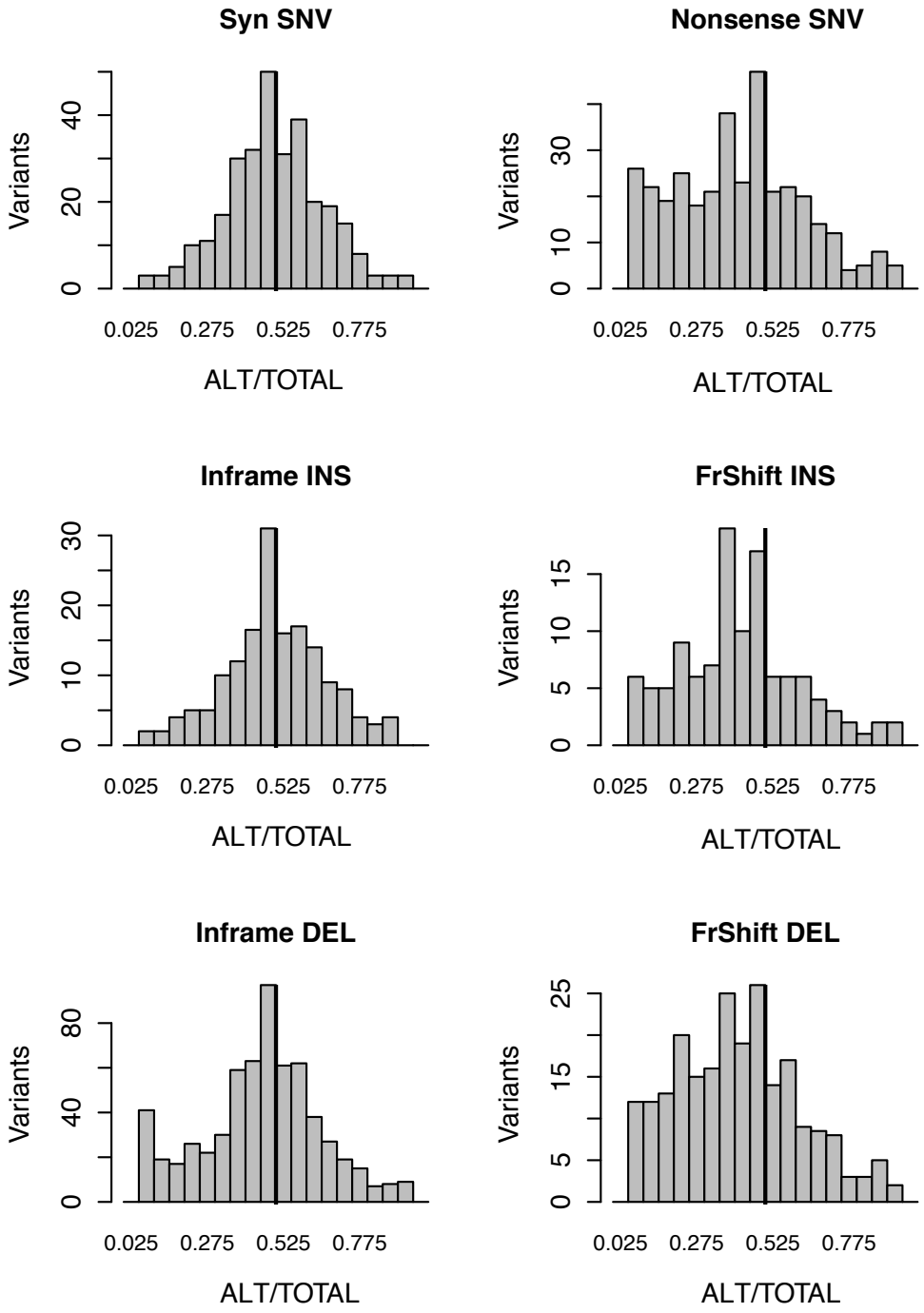


Figure S12: Allele-specific expression: allelic ratios of heterozygous genotypes with the predicted protein-truncating genetic variants (nonsense SNVs, frameshift insertions, frameshift deletions) shown on the right and the corresponding controls (synonymous SNVs, in-frame insertions, in-frame deletions) on the left. In order to capture an unbiased signal of NMD across variants, the data has been sampled to include only a single heterozygote individual per variant, and the histograms show the median distribution of 500 rounds of sampling.

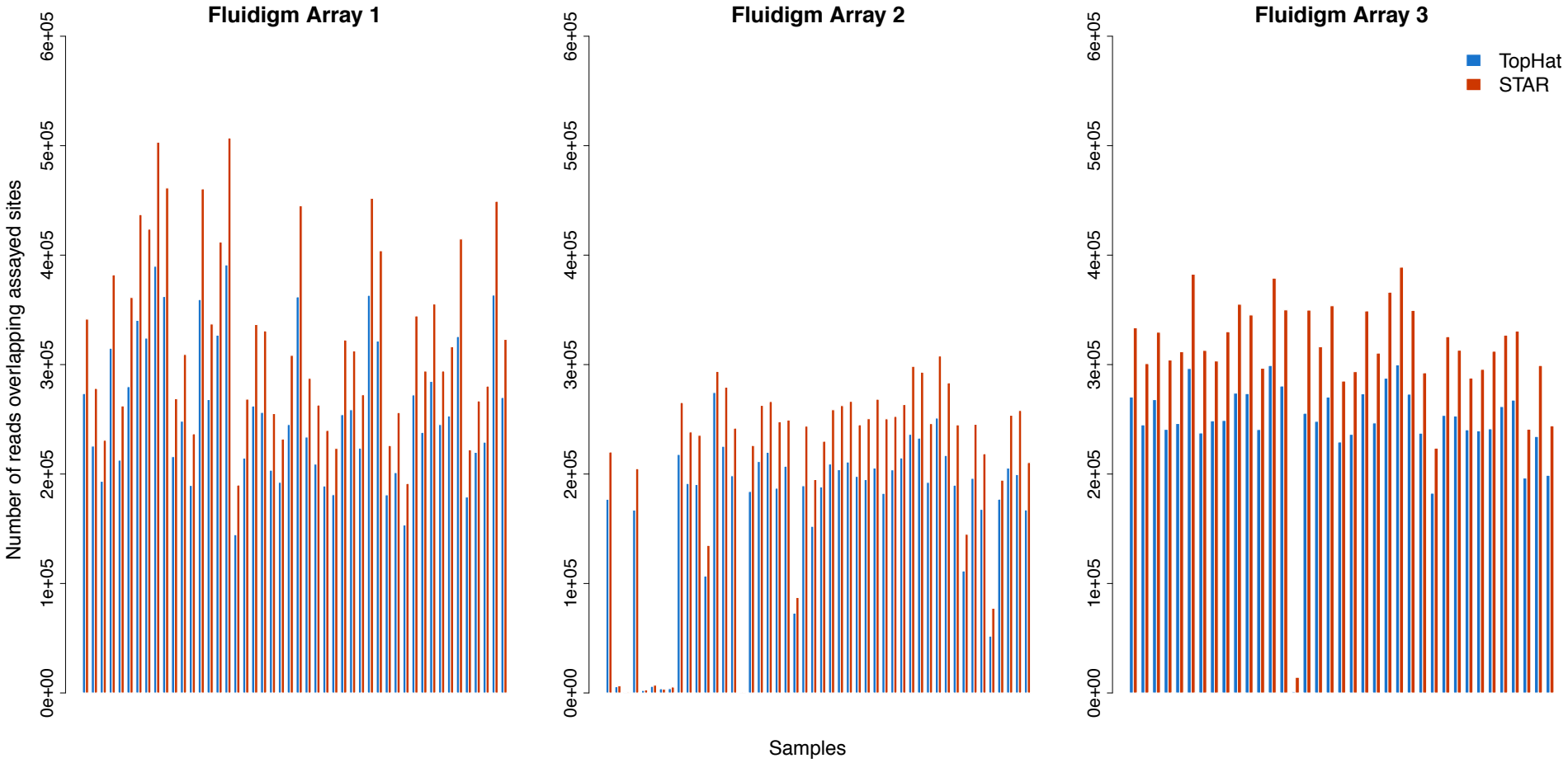


Figure S13: mmPCR-seq validation experiment: comparison of the number of reads overlapping targeted sites for different aligners. For both TopHat and STAR, we counted the number of reads that mapped to sites for which primers were designed. STAR mapped more reads to target sites across all samples.

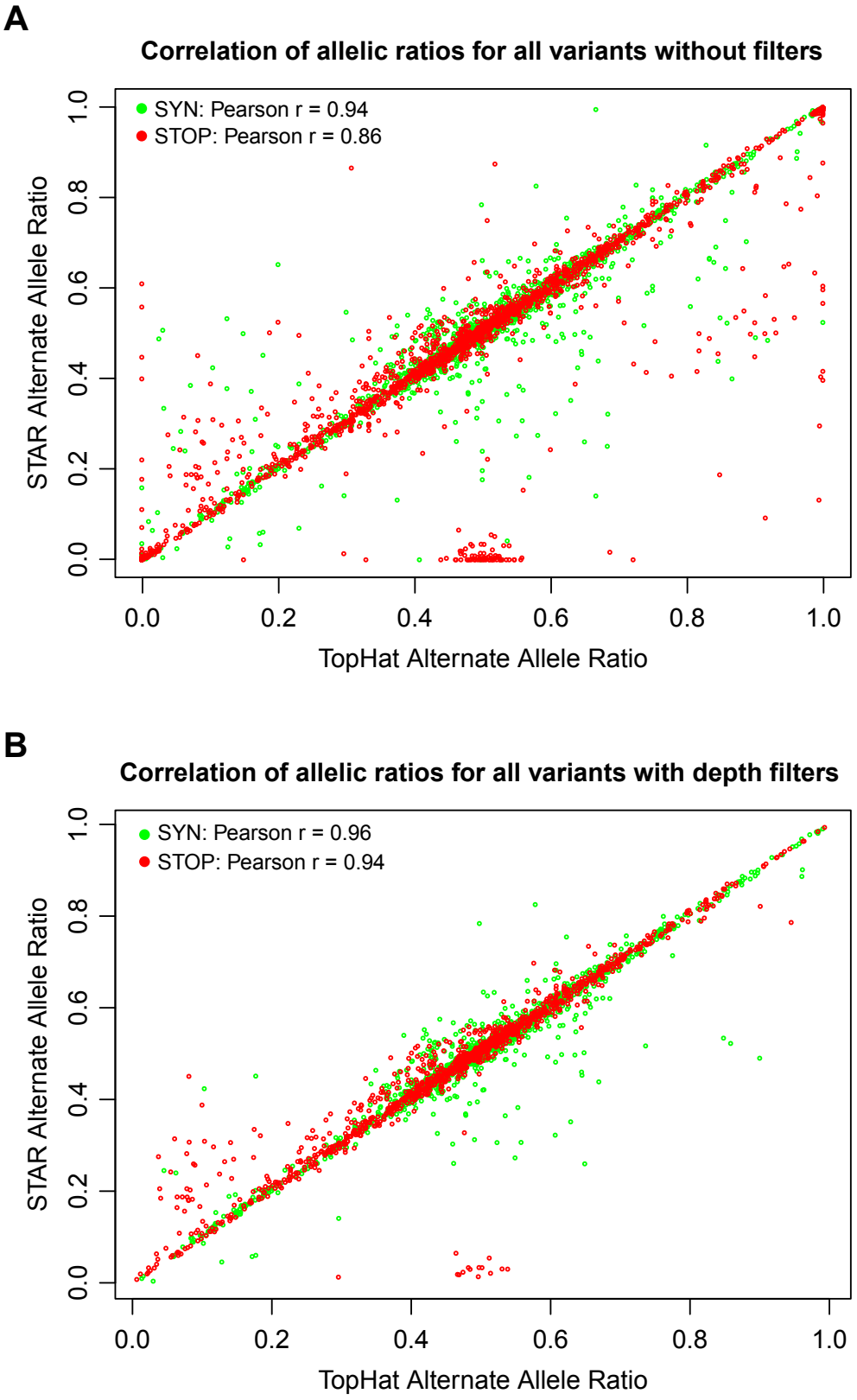


Figure S14: mmPCR-seq validation experiment: correlation of alternate allele ratios for different aligners. A. Correlation of allelic ratios for STAR and TopHat without depth filtering for synonymous (green) and nonsense variants (red). B. Correlation of allelic ratios after quality control filtering (total depth count > 150, ref allele count > 5, and non-ref allele count > 5). Except for one variant (rs1138349) that was tested in 9 samples, there is a high correlation of allelic ratios.

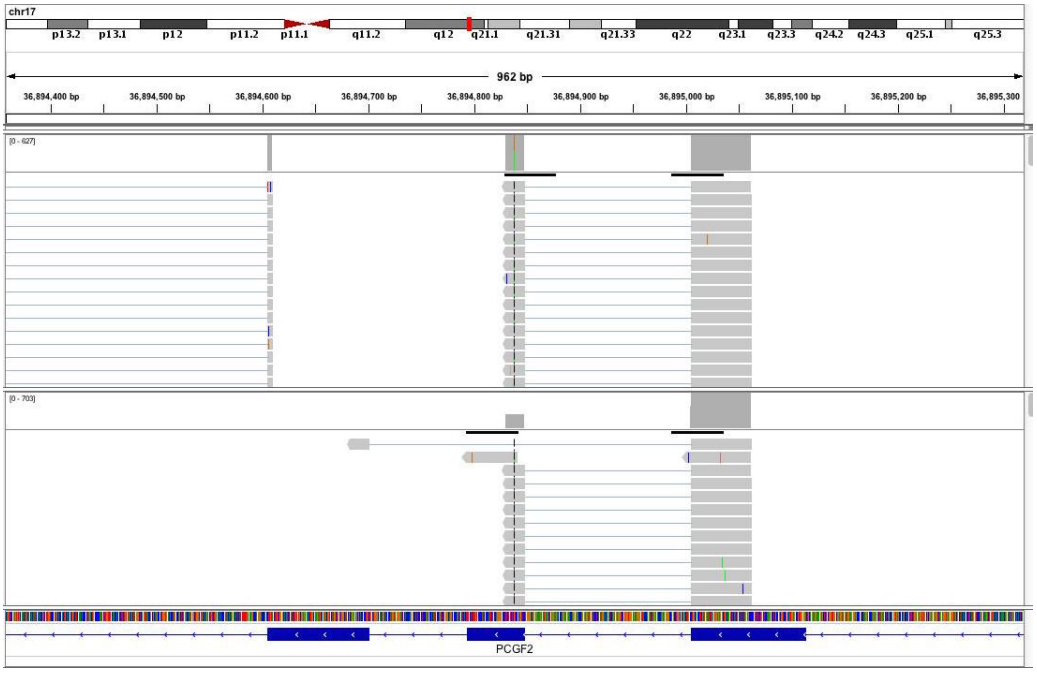
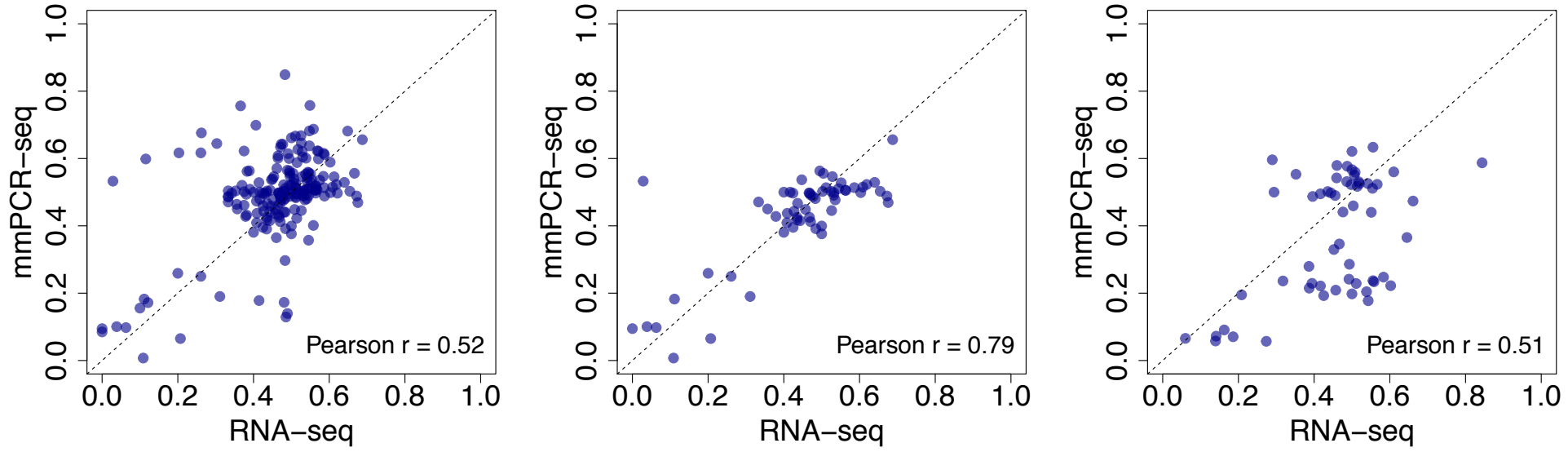


Figure S15: In Figure S14 a cluster of points is observed at $y = 0$ and $x = 0.5$, which corresponds to rs1138349. A possible explanation why STAR (bottom) fails to align one of the alleles is that the reads span three exons and two splice sites. STAR did a good job mapping reads spanning two exons and one splice junction. However, STAR failed to map reads that spanned all three exons and two splice junctions. Tophat alignment is shown at the top.



(a) All **(b)** Nonsense **(c)** Indels

Figure S16: mmPCR-seq validation experiment: correlation of alternate allele ratios measured from RNA-seq and mmPCR-seq. Only heterozygous SNV sites with at least 30 RNA-seq reads that passed mmPCR-seq quality control filtering (total depth count > 150, ref allele count > 5, and non-ref allele count > 5) were considered. The Pearson correlation is significant for a) all variants, b) nonsense variants, and c) indel ($P < 2.2 \times 10^{-16}$, $P = 7.3 \times 10^{-14}$, and $P = 4.4 \times 10^{-5}$).

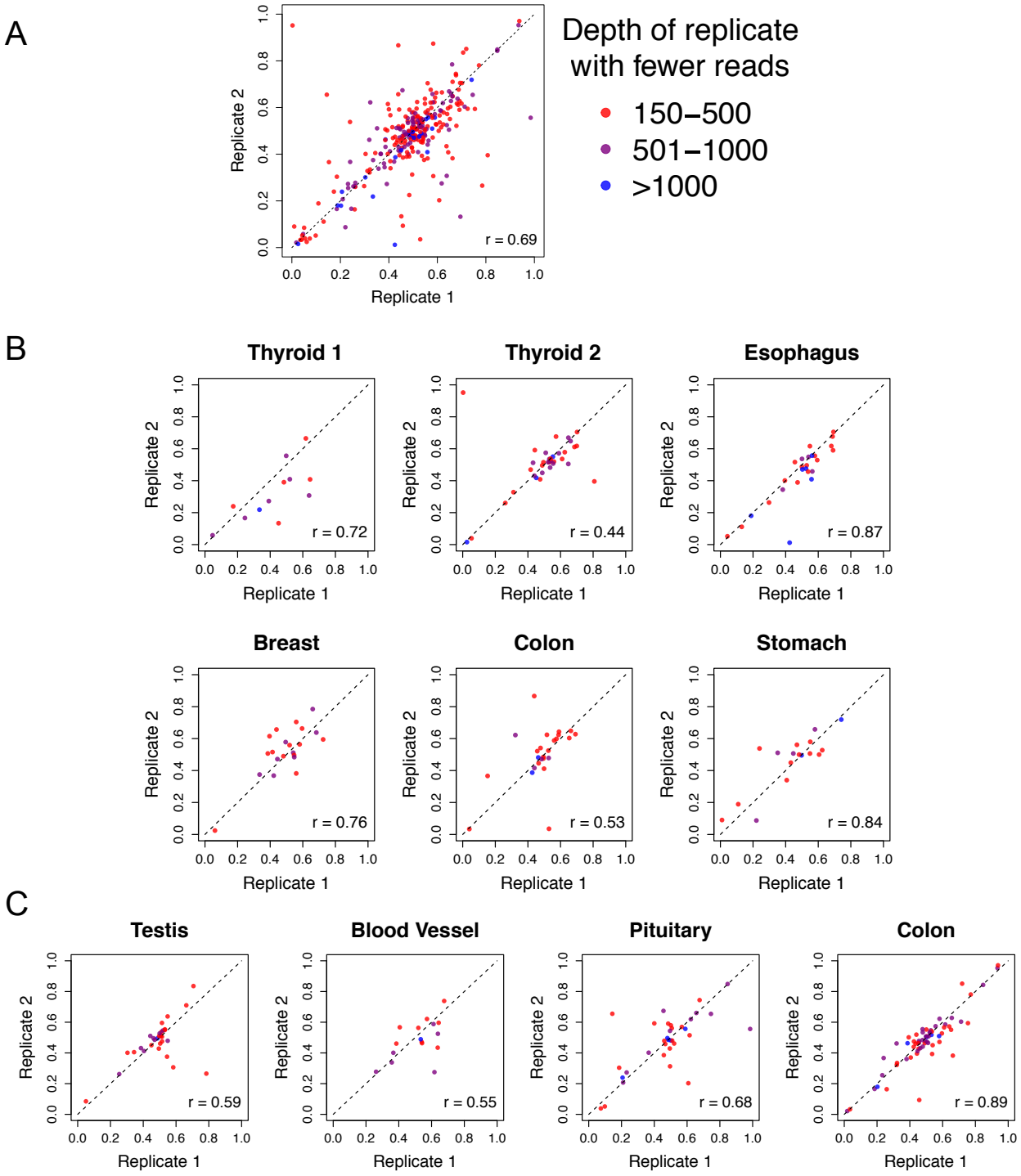


Figure S17: mmPCR-seq validation experiment: correlation of alternate allele ratio between technical replicates for all SNVs. We compared the alternate allele ratios at heterozygous SNV sites with at least 150 reads and at least five of each of the reference and alternate alleles using Pearson correlation. In each plot, sites are colored according to the number of reads in the replicate with lower depth at that site. a) All replicates pooled ($P < 2.2 \times 10^{-16}$). b) The six inter-array replicates with greater than ten sites that passed the depth threshold (thyroid, thyroid, esophagus, breast, colon, stomach $P = 0.008, 0.009, 8.1 \times 10^{-9}, 6.3 \times 10^{-5}, 0.006$, respectively). c) The inter-array replicates (testis, blood vessel, pituitary, colon $P = 5.2 \times 10^{-4}, 0.03, 1.1 \times 10^{-5}, 2.2 \times 10^{-16}$, respectively). The plots in b) and c) are labeled by the tissue of origin of the sample. The two thyroid samples come from different individuals.

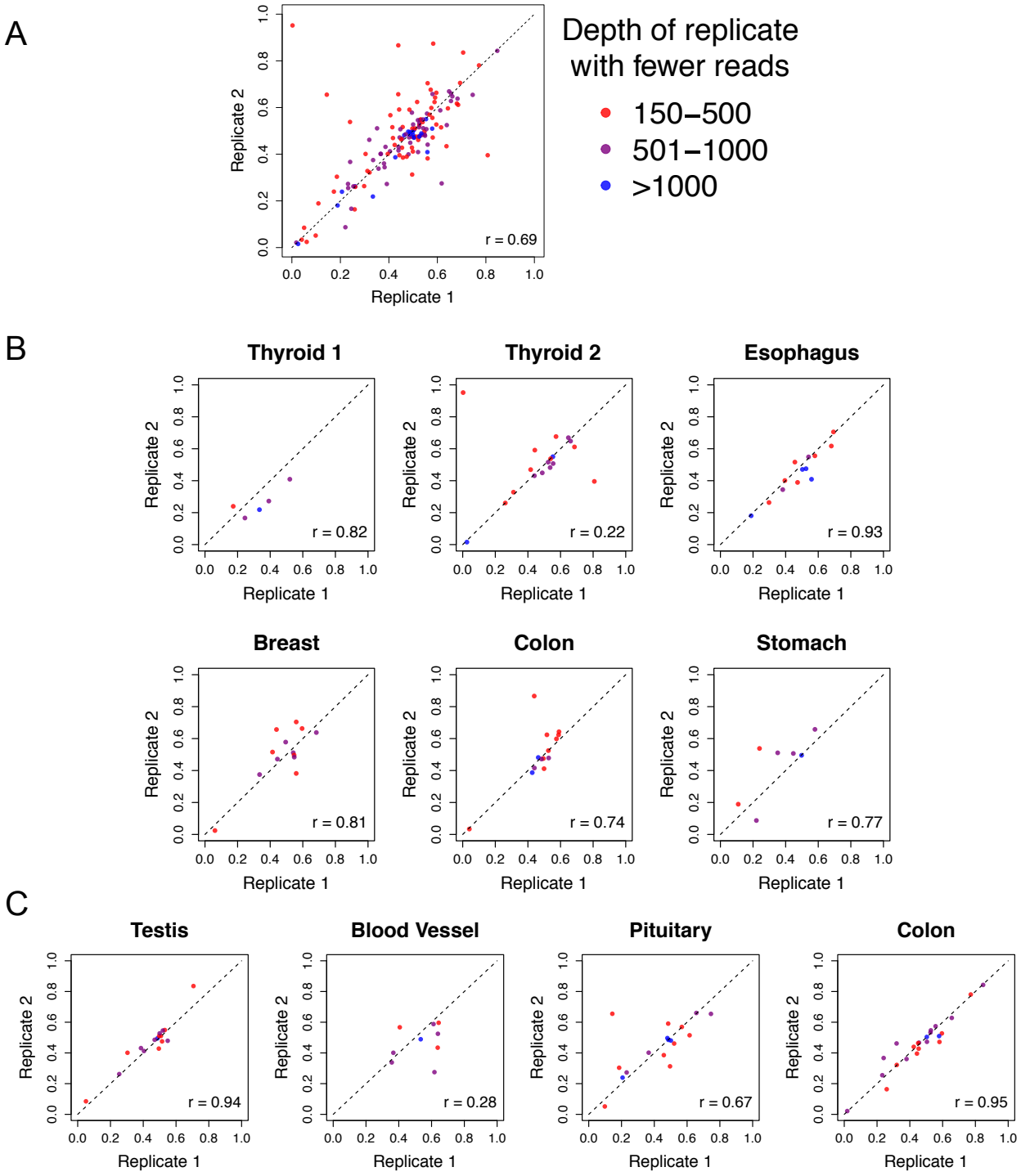


Figure S18: mmPCR-seq validation experiment: correlation of alternate allele ratio between technical replicates for nonsense SNVs only. We compared the alternate allele ratios at heterozygous SNV sites with at least 150 reads and at least five of each of the reference and alternate alleles using Pearson correlation. In each plot, sites are colored according to the number of reads in the replicate with lower depth at that site. The same replicate samples as in Figure S17 are shown. A. All replicates pooled ($P < 2.2 \times 10^{-16}$). B. Inter-array replicates (thyroid, thyroid, esophagus, breast, colon, stomach $P = 0.09, 0.38, 2.8 \times 10^{-6}, 7.4 \times 10^{-4}, 0.002, 0.05$, respectively). C. The inter-array replicates (testis, blood vessel, pituitary, colon $P = 3.8 \times 10^{-8}, 0.47, 0.004, 5.2 \times 10^{-12}$, respectively). The plots in b) and c) are labeled by the tissue of origin of the sample. The two thyroid samples come from different individuals.

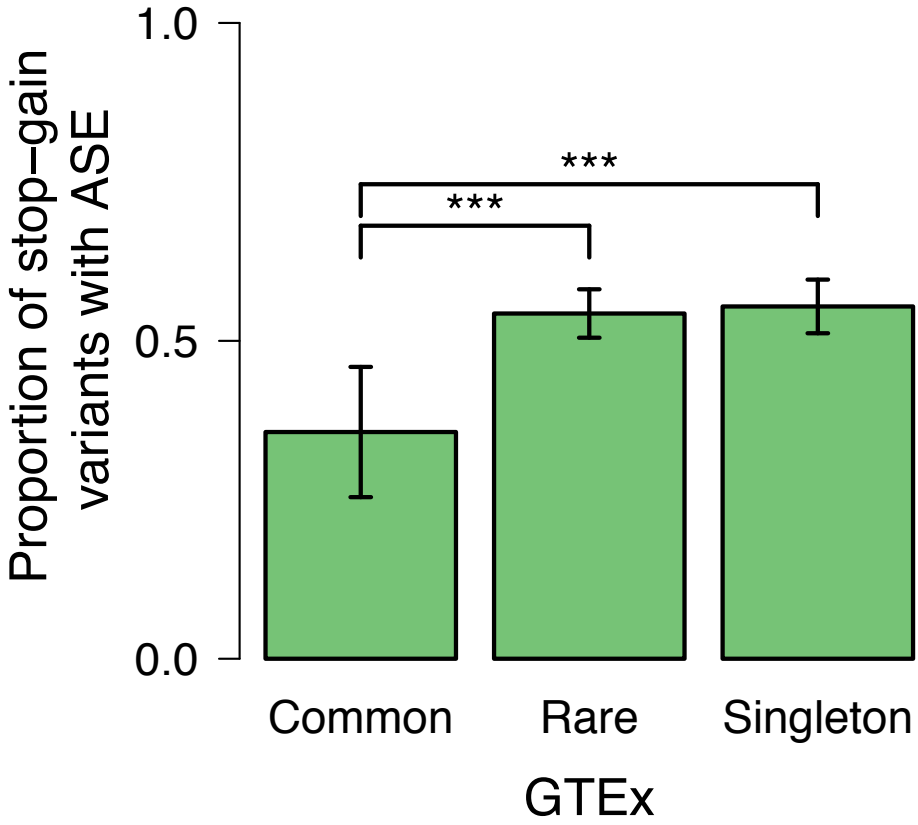


Figure S19: Proportion of nonsense variants with allele-specific expression effects in the GTEx data set in different frequency classes: common (MAF ≥ 0.05; $n = 84$), rare (MAF ≤ 0.01; $n = 657$), and singleton ($n = 532$) variants. We observed significantly higher proportion of allelic imbalance in rare and singleton nonsense variants (54.3%, 95% confidence interval (CI) 50.5% to 58.1%, and 55.4%, 95% CI $51.2 - 59.6\%$, respectively) compared to common nonsense variants (35.7%, 95% CI $25.4 - 45.9\%$).

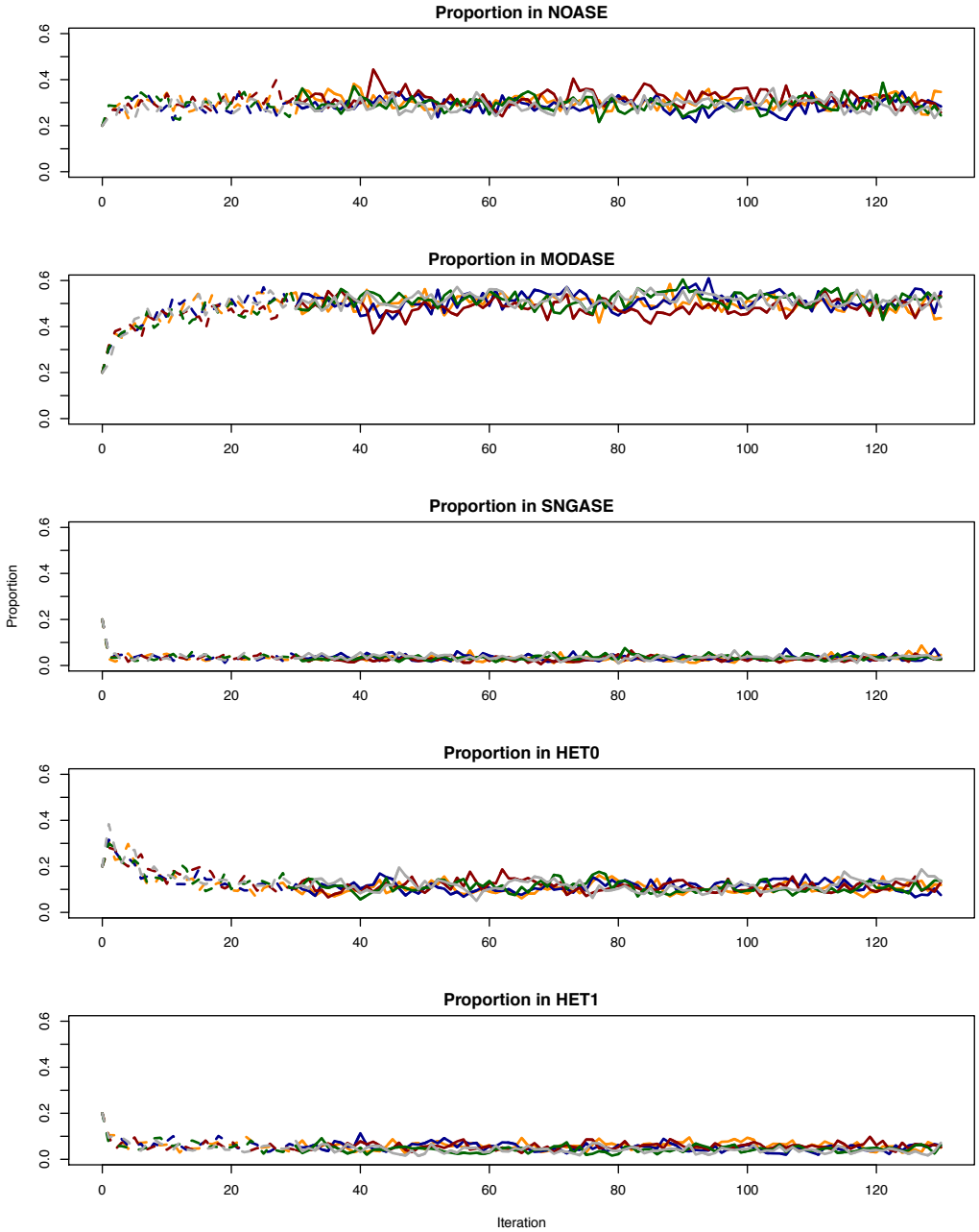


Figure S20: When applying MCMC algorithms to a data set it is customary to show the performance of the algorithm across all stages of the experiment (including the burn-in). We show that the GTM* algorithm generates stable proportion estimates for all of the five states studied: NOASE (no ASE effects across all tissues), MODASE (moderate ASE effects across all tissues), SNGASE (strong ASE across all tissues), HET0 (mixture of NOASE and/or MODASE, SNGASE) and HET1 (mixture of MODASE and SNGASE). For each state we demonstrate the proportion estimate during the burn-in stage of the experiment (30 iterations, dashed lines) and the state of the experiment used to obtain the global estimates (100 iterations, solid lines) reported in the manuscript. We highlight five different chains to show that the estimates are stable.

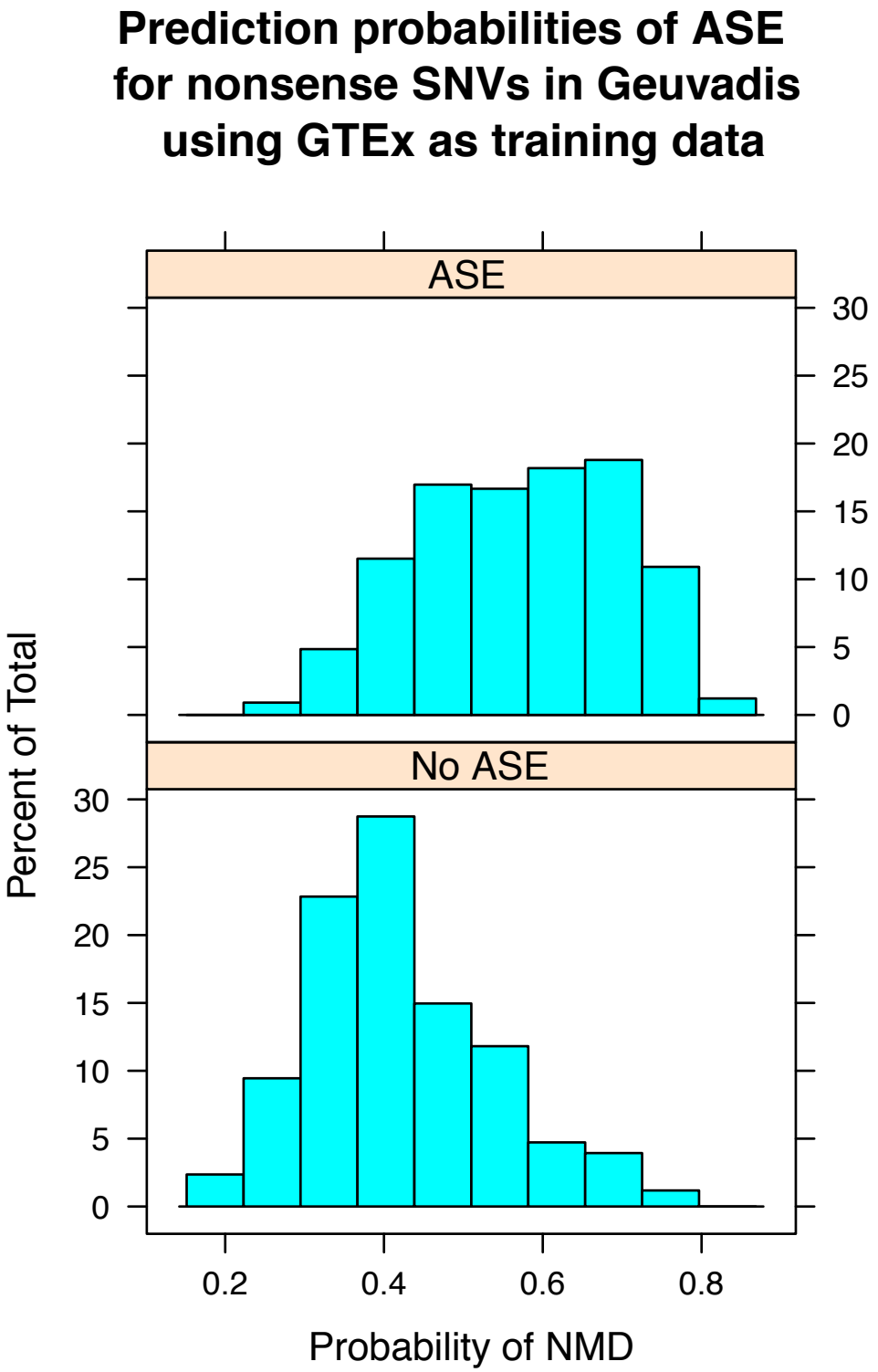
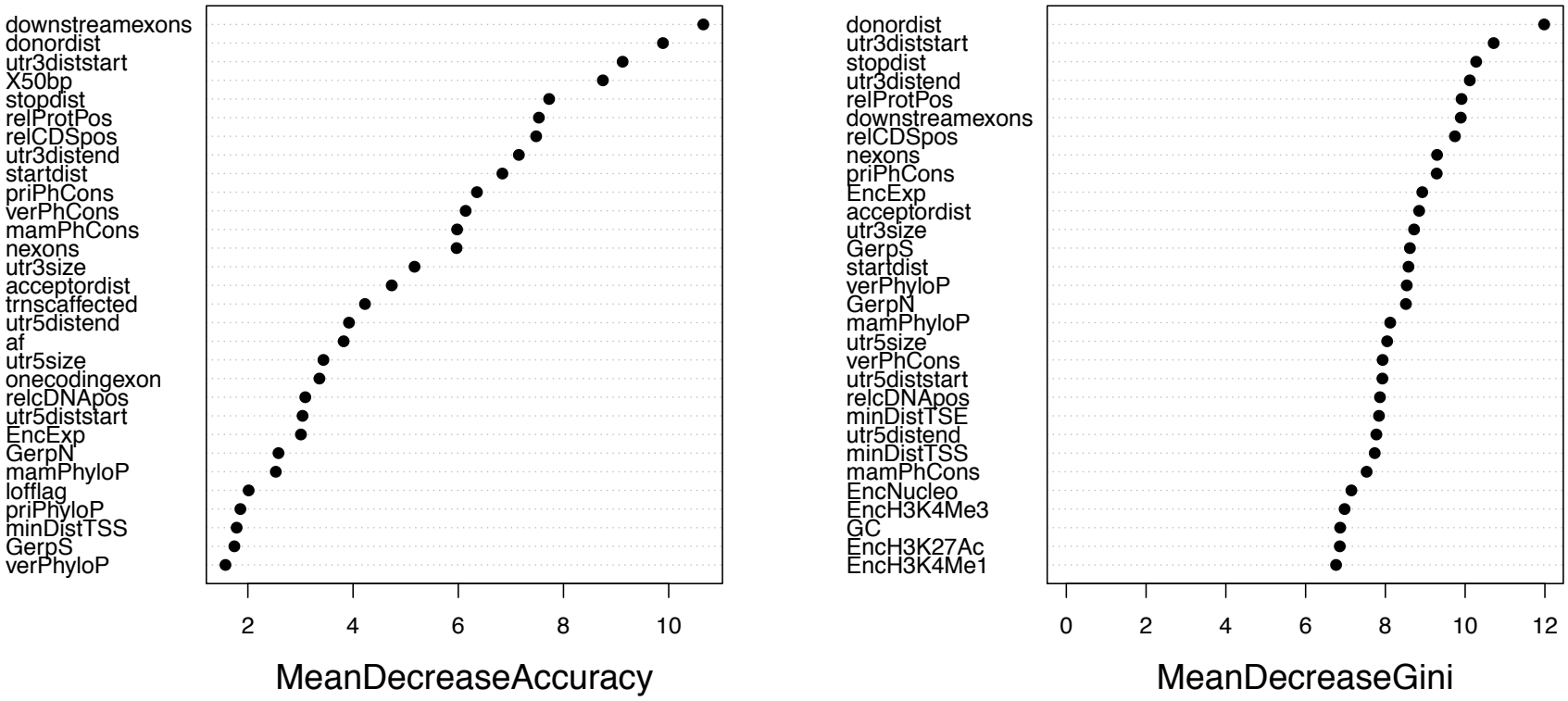


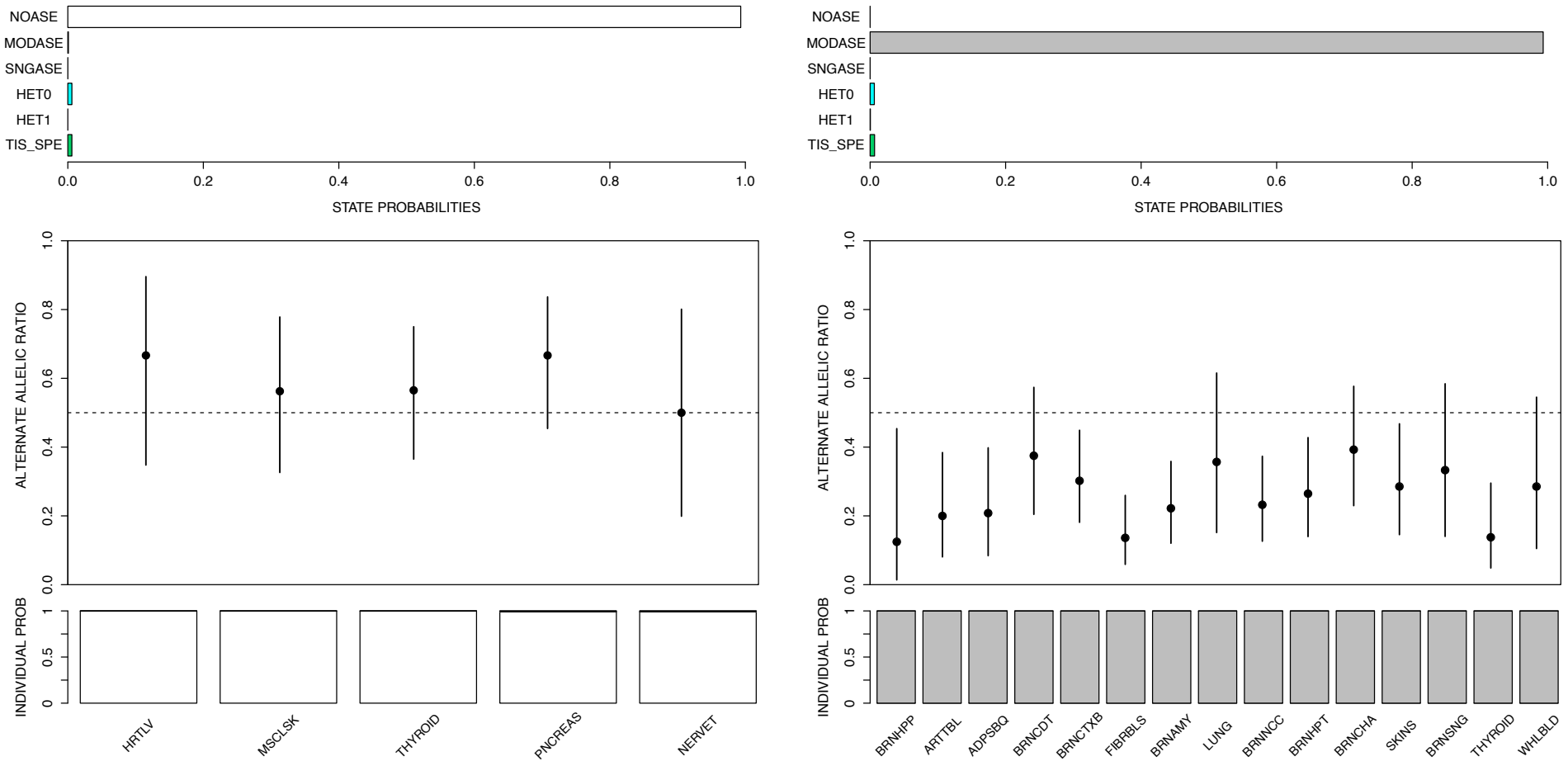
Figure S21: Insights into nonsense-mediated decay: modeling NMD with ASE outcome. When applying machine learning algorithms to a test data set that is independent from the training data set it is common practice to examine the prediction accuracy. To predict the outcome of the independent test data set (Geuvadis) we used the `predict.train` function (`caret` package) using the option `type = "prob"` to compute class probabilities. In this plot we show a histogram of the predicted probabilities of NMD assigned to nonsense SNVs in the Geuvadis data set. We show that for those variants that have no ASE the model predicted for 22.8% of variants to have probability $> .5$ of having ASE signal indicative of NMD. Conversely, it predicted for 77.2% of those variants to have probability $\leq .5$ of having ASE signal. Similarly, we show that for those variants that have some ASE the model predicted for 68.8% of variants to have probability $\geq .5$ of having ASE signal indicative of NMD. Conversely, it predicted for 31.2% of those variants that have some ASE to have probability $< .5$ of having ASE signal.



(a) Feature (variable) importance plot: mean decrease in accuracy

(b) Feature (variable) importance plot: mean decrease in gini

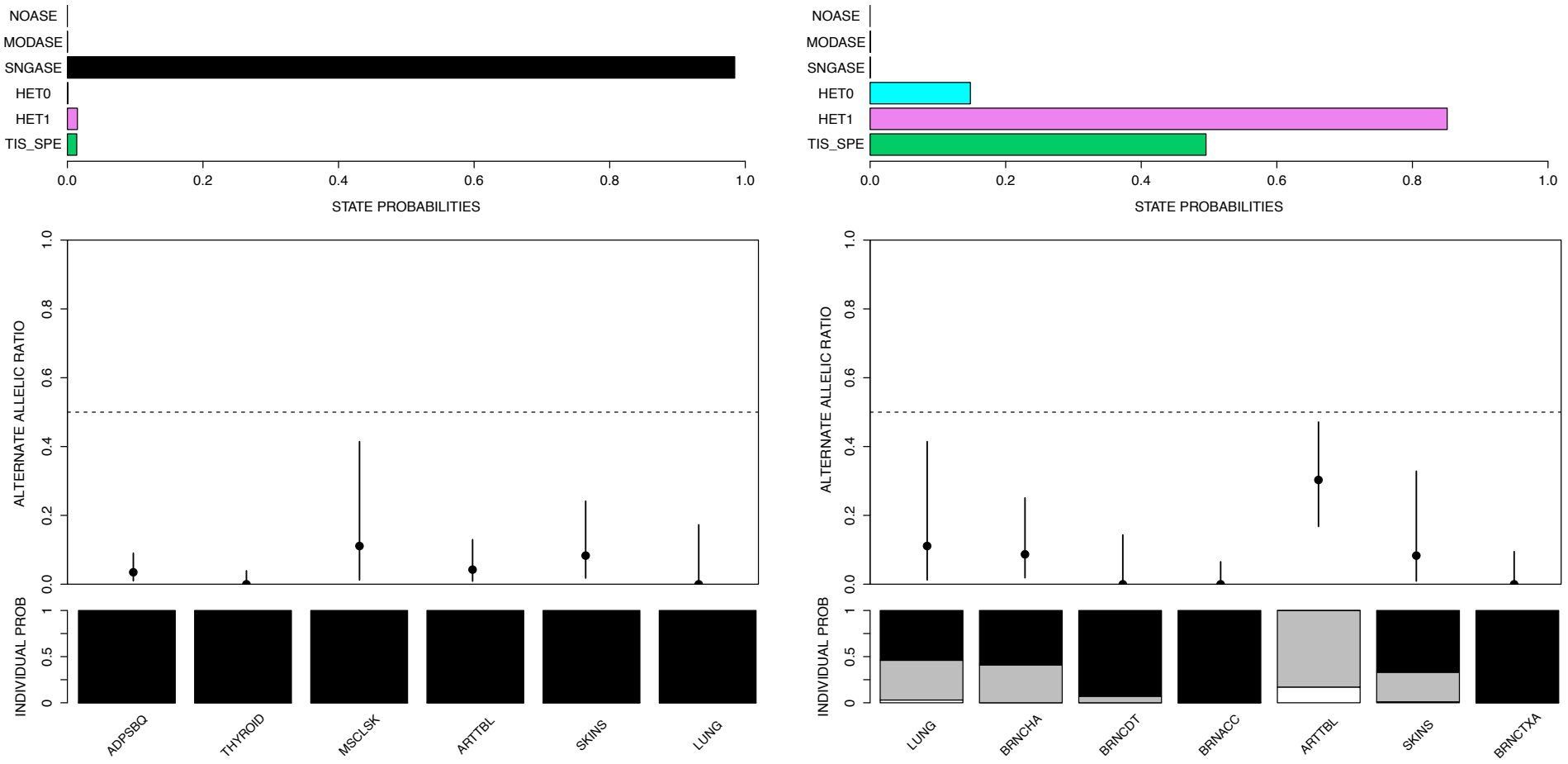
Figure S22: Insights into nonsense-mediated decay: feature (variable) importance plots for the random forest algorithm. For each feature in the training exercise the feature importance plot tells the user how important that feature is in classifying the data. The top 30 features are represented on the graphs and an estimate of their importance is given on the x-axis. In the training exercise we have an 80:20 percent split, i.e. treating 80% of the GTEx data set as the training set and 20% as the test set (commonly referred to as the out of bag observations). a) By contrasting the out of bag predictions with the known outcomes, we arrive at an estimate of the prediction error rate. For each feature, we compute the mean decrease in accuracy by comparing the prediction error rate to the case when the values of the feature are permuted. b) A higher mean decrease in Gini means that the feature plays a greater role in partitioning the data into the defined classes (some ASE or no ASE). For a description and a list of the 38 sequence and genomic features used see table S5.



(a) NO ASE

(b) Moderate ASE across all tissues

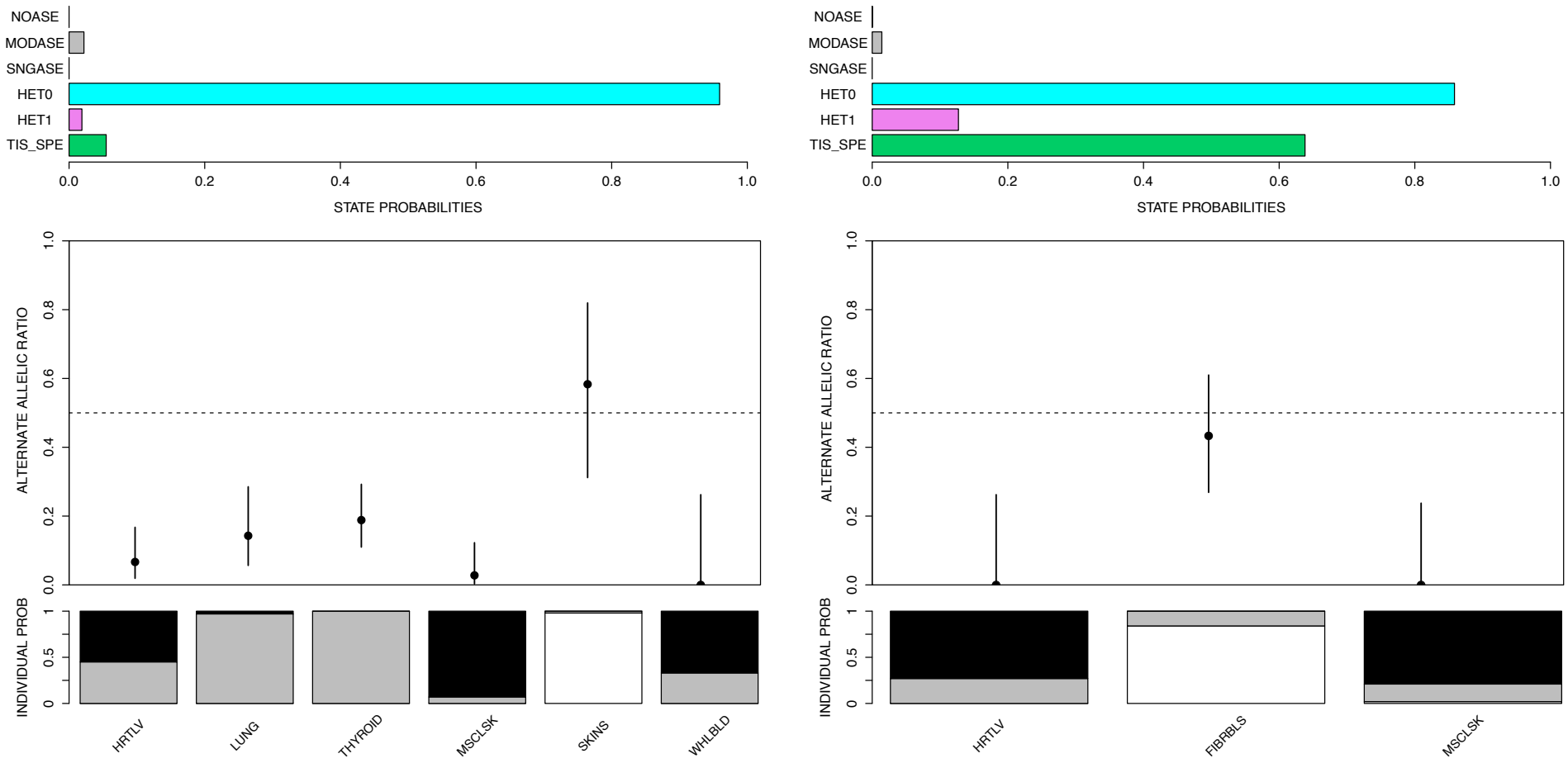
Figure S23: ASE classification examples: no ASE and moderate ASE across all tissues. The plots shown represent several properties of the data: (top) posterior probabilities of the variant classification states across all tissues including NOASE (no ASE effects across all tissues), MODASE (moderate ASE effects across all tissues), SNGASE (strong ASE effects across all tissues), HET0 (heterogeneous effects with some tissues having no ASE and others some form of ASE), HET1 (heterogeneous effects with some tissues having moderate ASE effects others strong ASE effects), TIS_SPE (tissue specific effect); (center) shows the alternate allelic ratio (maximum likelihood estimate and 95% confidence interval) per tissue; (bottom) INDIVIDUAL PROB representing the individual tissue posterior probability of no ASE (white), moderate ASE (gray), or strong ASE (black). a) An example of a PTV, p.S88X (rs41296182) in the gene *TRIM45* (tripartite motif containing 4), classified as having no ASE effects across all studied tissues (posterior probability for the NOASE state = 0.99). b) An example of a PTV, p.Q59X (rs121908176) in the gene *BBS2* (Bardet-Biedl syndrome 2), classified as having moderate ASE effects across all studied tissues (posterior probability for the MODASE state = 0.99).



(a) Strong ASE across all tissues

(b) Mixture of moderate and strong ASE

Figure S24: ASE classification examples: strong ASE across all tissues and mixture of moderate and strong ASE. The plots shown represent several properties of the data: (top) posterior probabilities of the variant classification states across all tissues including NOASE (no ASE effects across all tissues), MODASE (moderate ASE effects across all tissues), SNGASE (strong ASE effects across all tissues), HET0 (heterogeneous effects with some tissues having no ASE and others some form of ASE), HET1 (heterogeneous effects with some tissues having moderate ASE effects others strong ASE effects), TIS_SPE (tissue specific effect); (center) shows the alternate allelic ratio (maximum likelihood estimate and 95% confidence interval) per tissue; (bottom) INDIVIDUAL PROB representing the individual tissue posterior probability of no ASE (white), moderate ASE (gray), or strong ASE (black). a) An example of a PTV, p.R388X (snp_2_201476115) in the gene *AOX1* (aldehyde oxidase 1), classified as having strong ASE effects across all studied tissues (posterior probability for the SNGASE state $= 0.98$). b) An example of a PTV, p.E318X (snp_6_86256830) in the gene *SNX14* (sorting nexin 14), classified as having a mixture of moderate and strong ASE effects (posterior probability for the HET1 state $= 0.87$).



(a) Mixture of no ASE and ASE

(b) Tissue-specific ASE

Figure S25: ASE classification examples: mixture of no ASE and ASE effect and tissue-specific ASE. The plots shown represent several properties of the data: (top) posterior probabilities of the variant classification states across all tissues including NOASE (no ASE effects across all tissues), MODASE (moderate ASE effects across all tissues), SNGASE (strong ASE effects across all tissues), HET0 (heterogeneous effects with some tissues having no ASE and others some form of ASE), HET1 (heterogeneous effects with some tissues having moderate ASE effects others strong ASE effects), TIS_SPE (tissue specific effect); (center) shows the alternate allelic ratio (maximum likelihood estimate and 95% confidence interval) per tissue; (bottom) INDIVIDUAL PROB representing the individual tissue posterior probability of no ASE (white), moderate ASE (gray), or strong ASE (black). a) An example of a PTV, p.Q776X (rs149244943) in the gene *PHKB* (phosphorylase kinase, beta), classified as having a mixture of no ASE and ASE effects (posterior probability for the HET0 state $= 0.96$). b) An example of a PTV, p.Q66X (snp_14_7817080) in the gene *ALKBH1* (alkB, alkylation repair homolog 1), classified as having a tissue-specific ASE effect (posterior probability for the TIS_SPE state $= 0.64$).

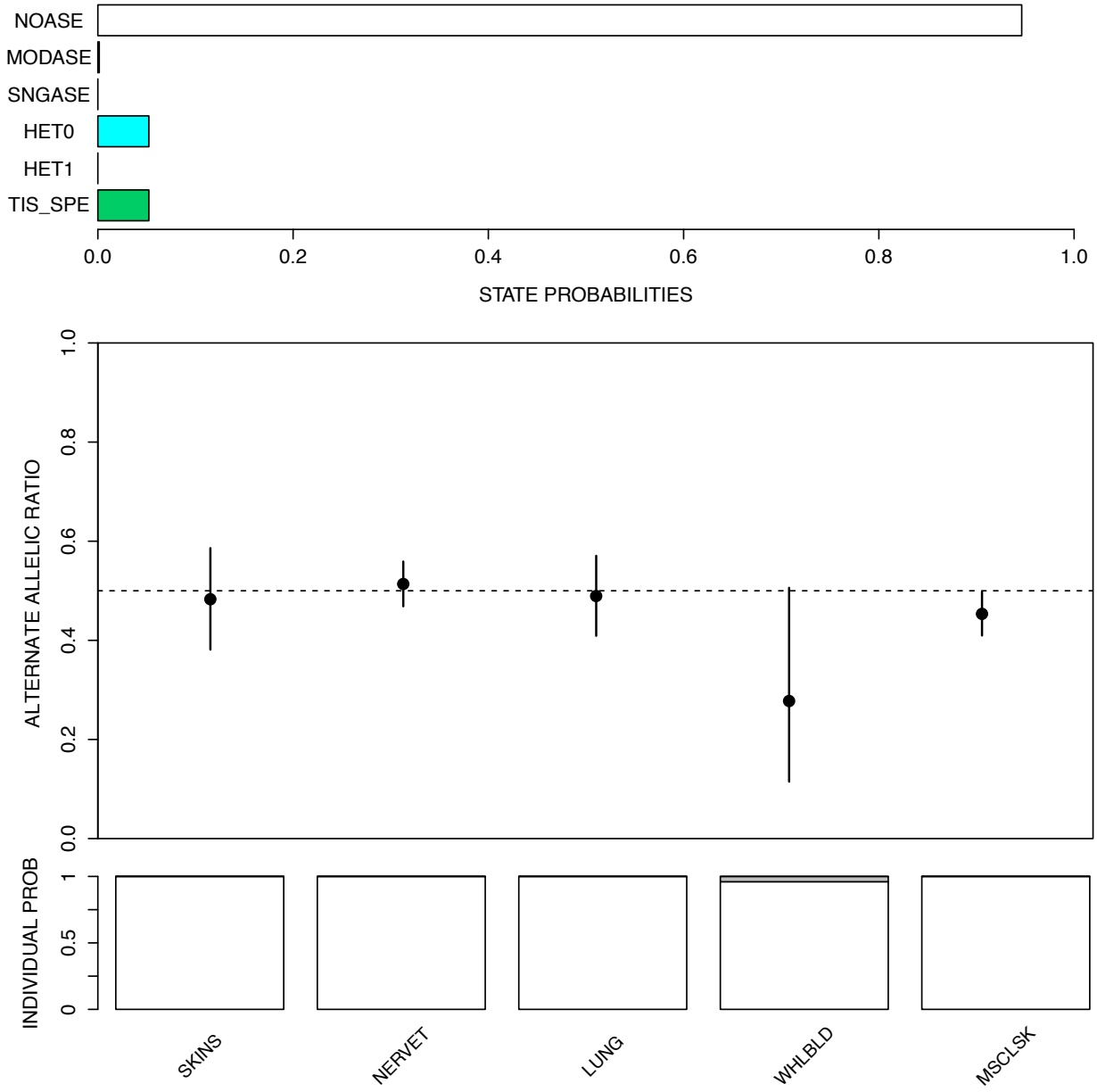
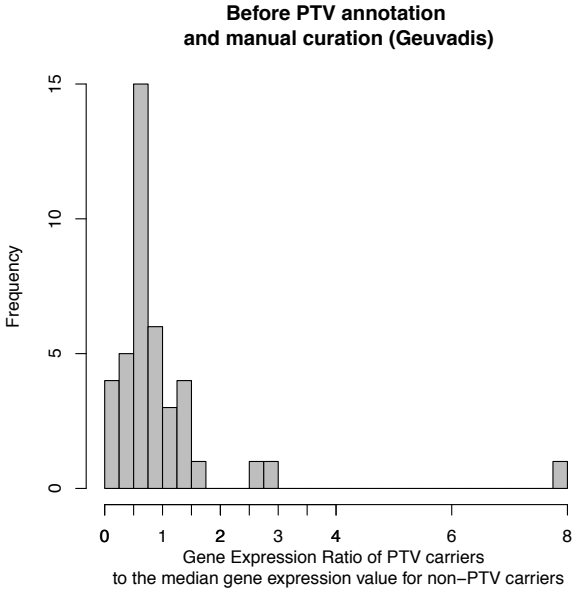
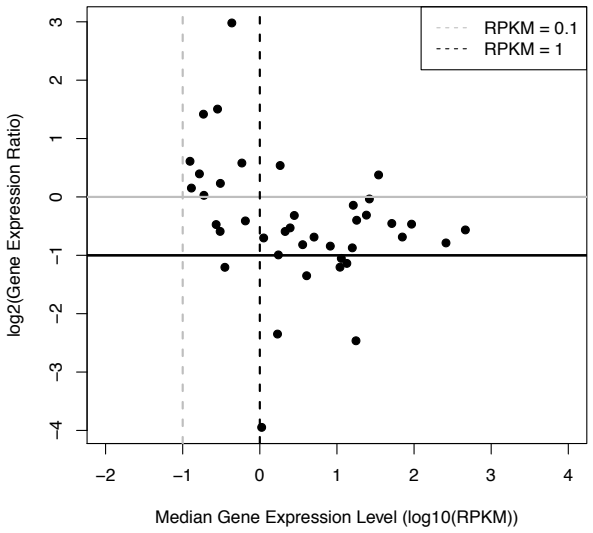


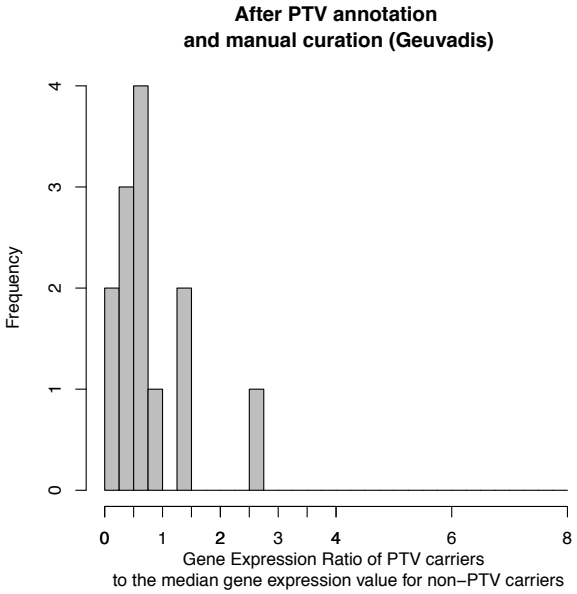
Figure S26: ASE data for p.S474X (rs328) in the gene *LPL* (lipoprotein lipase). The variant is classified as having no ASE across all tissues (posterior probability for the NOASE state > 0.99) in the RNA-seq data set, supporting the observation that transcripts with the mutation are retained. This is consistent with reports of a truncated protein with a gain-of-function mutation and suggests that such proteins are likely present across all tissues.



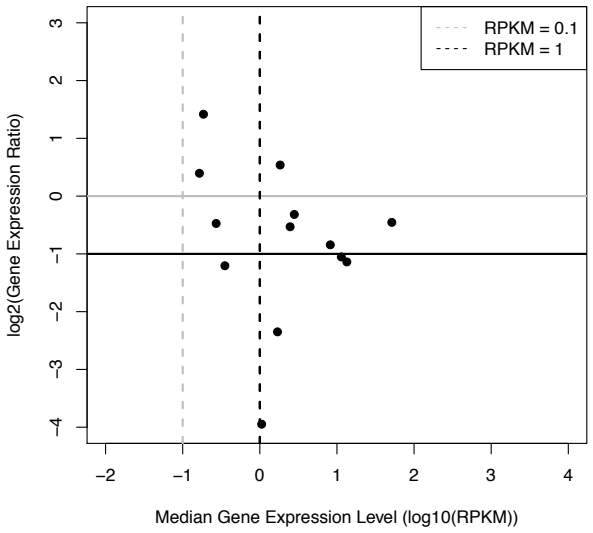
(a) Histogram of gene expression ratios



(b) Scatter plot of the median gene expression value and gene expression ratio

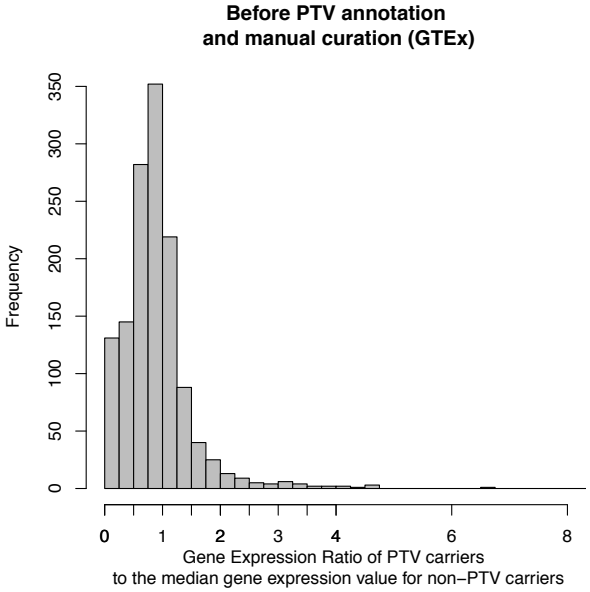


(c) Histogram of gene expression ratios

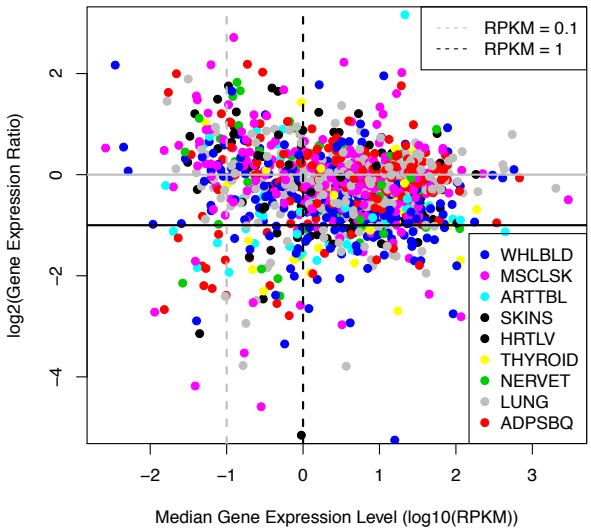


(d) Scatter plot of the median gene expression value and gene expression ratio

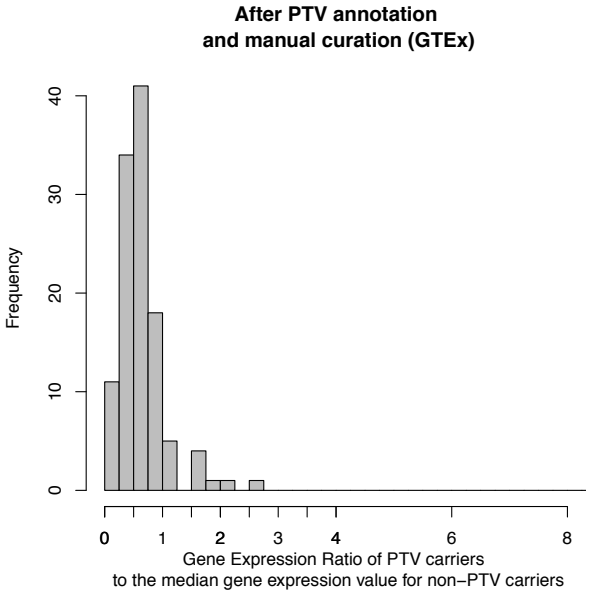
Figure S27: Insights into dosage compensation: examining gene expression ratios measured by comparing the gene expression value of the deleted gene(s) in PTV carriers to the median gene expression value of non-PTV carriers. In a) and c) we show the histogram of the gene expression ratios in the Geuvadis data set before ($n = 27$ large deletions) and after ($n = 8$ large deletions) PTV annotation and manual curation filtering, respectively. In b) and d) we show a scatter plot of the median gene expression across all individuals (x-axis) and the gene expression ratio (y-axis) before and after PTV annotation and manual curation filtering, respectively.



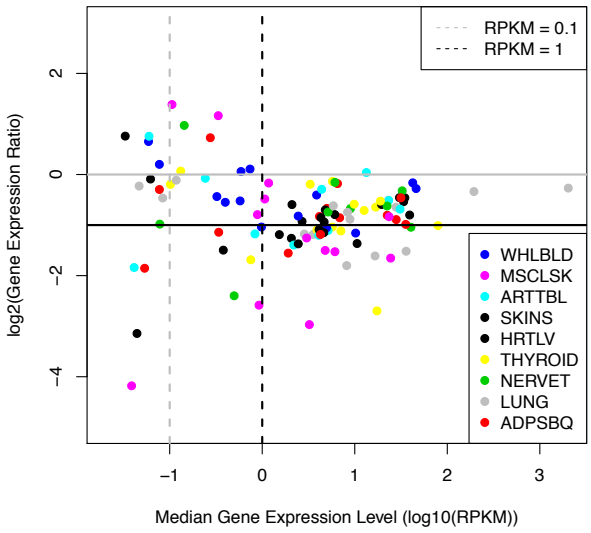
(a) Histogram of gene expression ratios



(b) Scatter plot of the median gene expression value and gene expression ratio

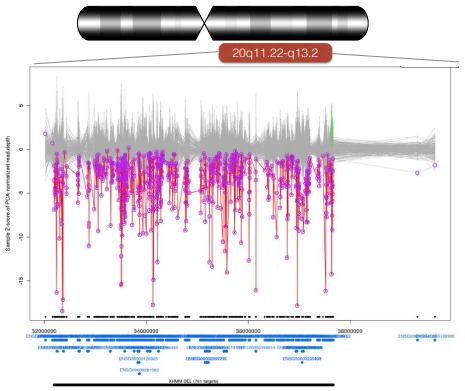


(c) Histogram of gene expression ratios

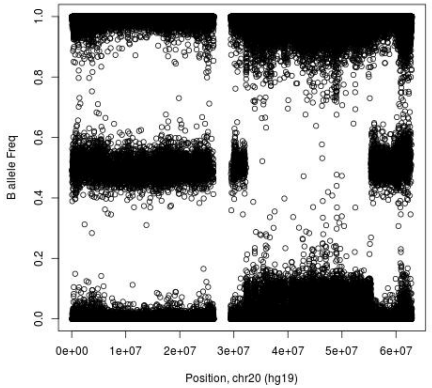


(d) Scatter plot of the median gene expression value and gene expression ratio

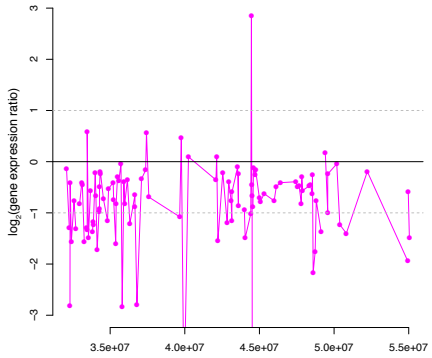
Figure S28: Insights into dosage compensation: examining gene expression ratios measured by comparing the gene expression value of the deleted gene(s) in PTV carriers to the median gene expression value of non-PTV carriers. In a) and c) we show the histogram of the gene expression ratios in the GTEx data set before ($n = 65$ large deletions) and after ($n = 3$ large deletions) PTV annotation and manual curation filtering, respectively. In b) and d) we show a scatter plot of the median gene expression across all individuals (x-axis) and the gene expression ratio (y-axis) before and after PTV annotation and manual curation filtering, respectively.



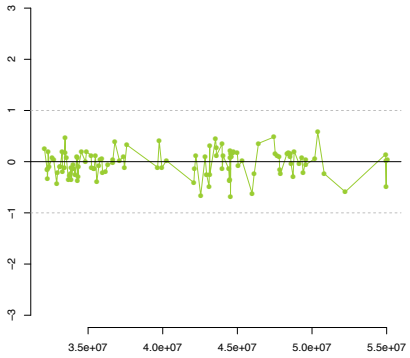
(a) XHMM 20Mb deletion



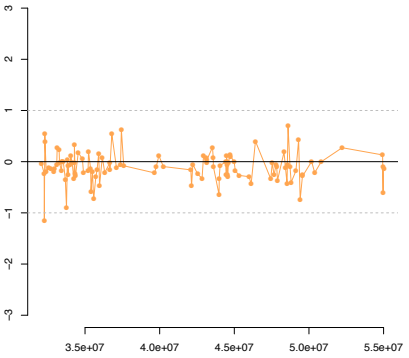
(b) XHMM 20Mb deletion (omni 2.5M array data)



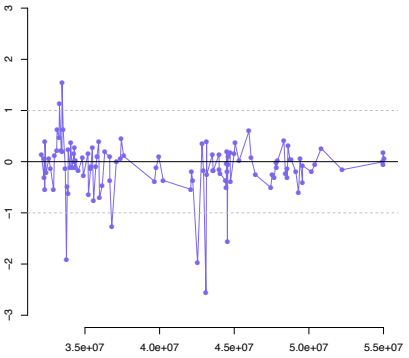
(c) Blood (WHLBLD)



(d) Lung (LUNG)

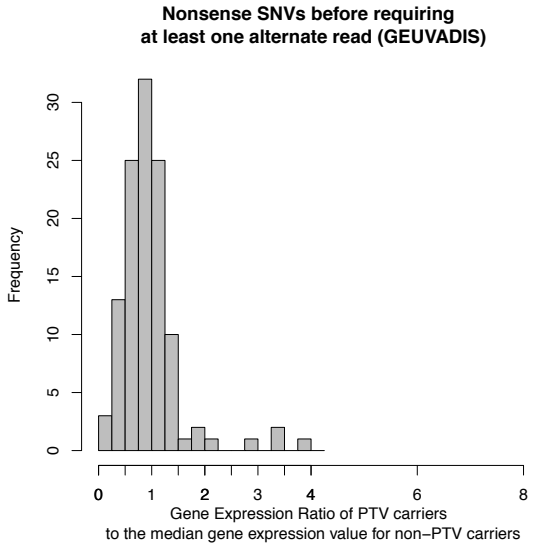


(e) Adipose Subcutaneous (ADPSBQ)

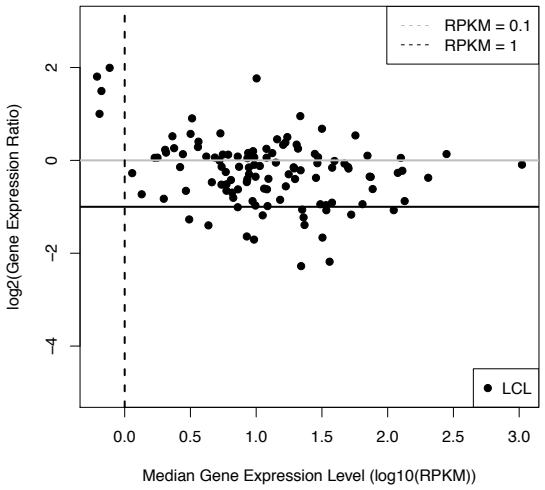


(f) Muscle Skeletal (MSCLSK)

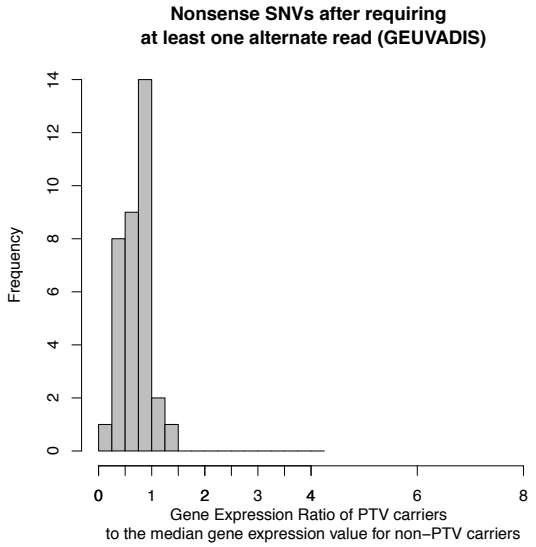
Figure S29: Insights into dosage compensation: impact of somatic variants. a) In the GTEx exome sequencing data set we detected a large deletion event in a single individual that spanned over 20 megabases (Mb). We show in red the sample Z-score of PCA normalized read depth for the sample containing the 20Mb event and in gray for all other samples. b) Using data from an Omni 2.5M SNP array, we were able to confirm the presence of a large 20Mb somatic deletion. Across the predicted deletion interval, there are essentially no B allele frequency (BAF) values corresponding to a typical heterozygous SNP (expected mean BAF $= 0.5$). Within this same region, there is an increased density of points around BAF $= 0.9$ and BAF $= 0.1$, reflecting the fact that we have detected a mosaic deletion. Somatic mosaicism in blood was inferred by the frequency of heterozygous genotypes and confirmed by the patterns of $\log_2$ gene expression ratios (individual carrier gene expression value/median non-PTV expression values, y-axis) across the 20Mb region in c) whole blood (mosaic deletion is present), d) lung (mosaic deletion is absent), e) subcutaneous adipose (mosaic deletion is absent), and f) skeletal muscle (mosaic deletion is absent). When studying dosage compensation the patterns of no difference in gene expression across all the genes in the 20Mb deletion in lung, adipose, and muscle could be mistaken as evidence of gene dosage compensation.



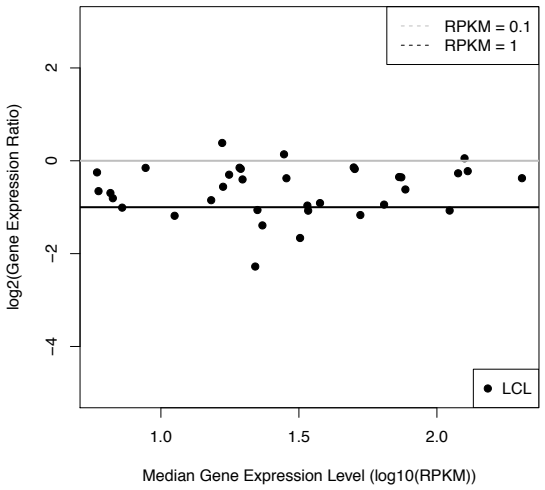
(a) Histogram of gene expression ratios



(b) Scatter plot of the median gene expression value and gene expression ratio

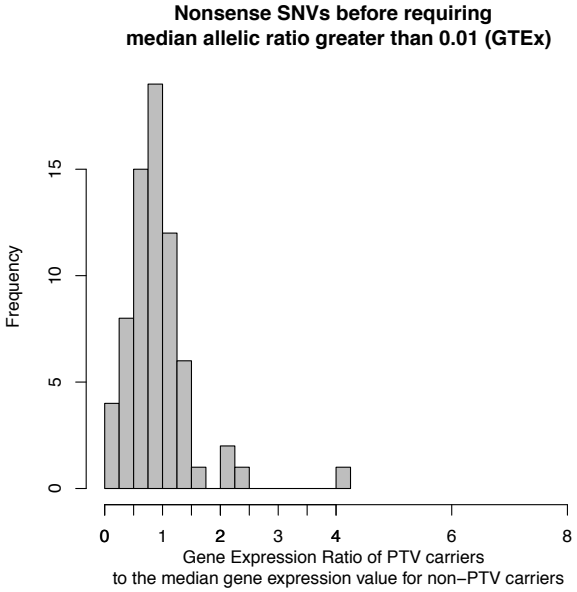


(c) Histogram of gene expression ratios

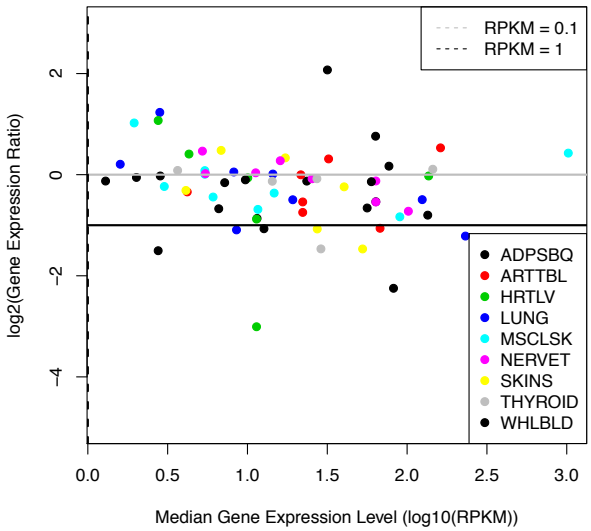


(d) Scatter plot of the median gene expression value and gene expression ratio

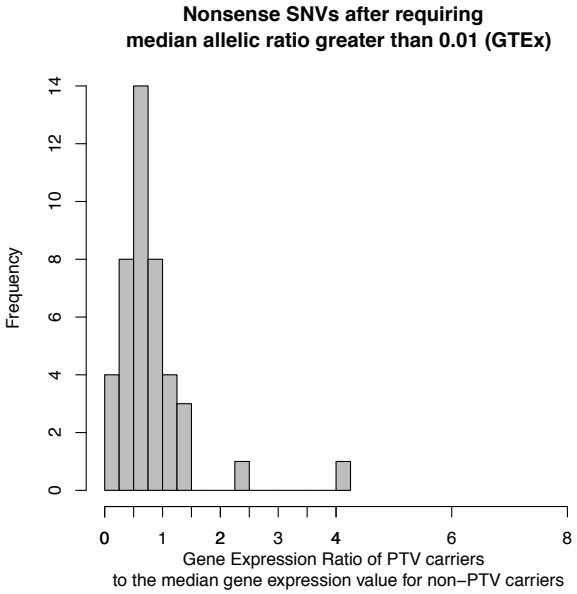
Figure S30: Insights into dosage compensation: examining gene expression ratios measured by comparing the gene expression value of the PTV carrier to the median gene expression value of non-PTV carriers. In a) and c) we show the histogram of the gene expression ratios in the GEUVADIS data set before ($n = 116$ nonsense SNVs) and after ($n = 35$ nonsense SNVs) requiring at least one alternate read in the ASE data set, respectively. In b) and d) we show a scatter plot of the median gene expression across all individuals (x-axis) and the gene expression ratio (y-axis) before and after filtering, respectively.



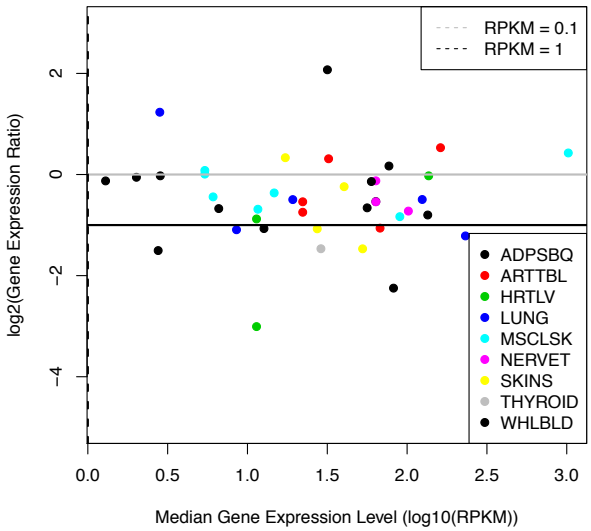
(a) Histogram of gene expression ratios



(b) Scatter plot of the median gene expression value and gene expression ratio



(c) Histogram of gene expression ratios



(d) Scatter plot of the median gene expression value and gene expression ratio

Figure S31: Insights into dosage compensation: examining gene expression ratios measured by comparing the gene expression value of the PTV carrier to the median gene expression value of non-PTV carriers. In a) and c) we show the histogram of the gene expression ratios in the GTEx data set before ($n = 25$ nonsense SNVs) and after ($n = 18$ nonsense SNVs) requiring at least one alternate read in the ASE data set, respectively. In b) and d) we show a scatter plot of the median gene expression across all individuals (x-axis) and the gene expression ratio (y-axis) before and after filtering, respectively.

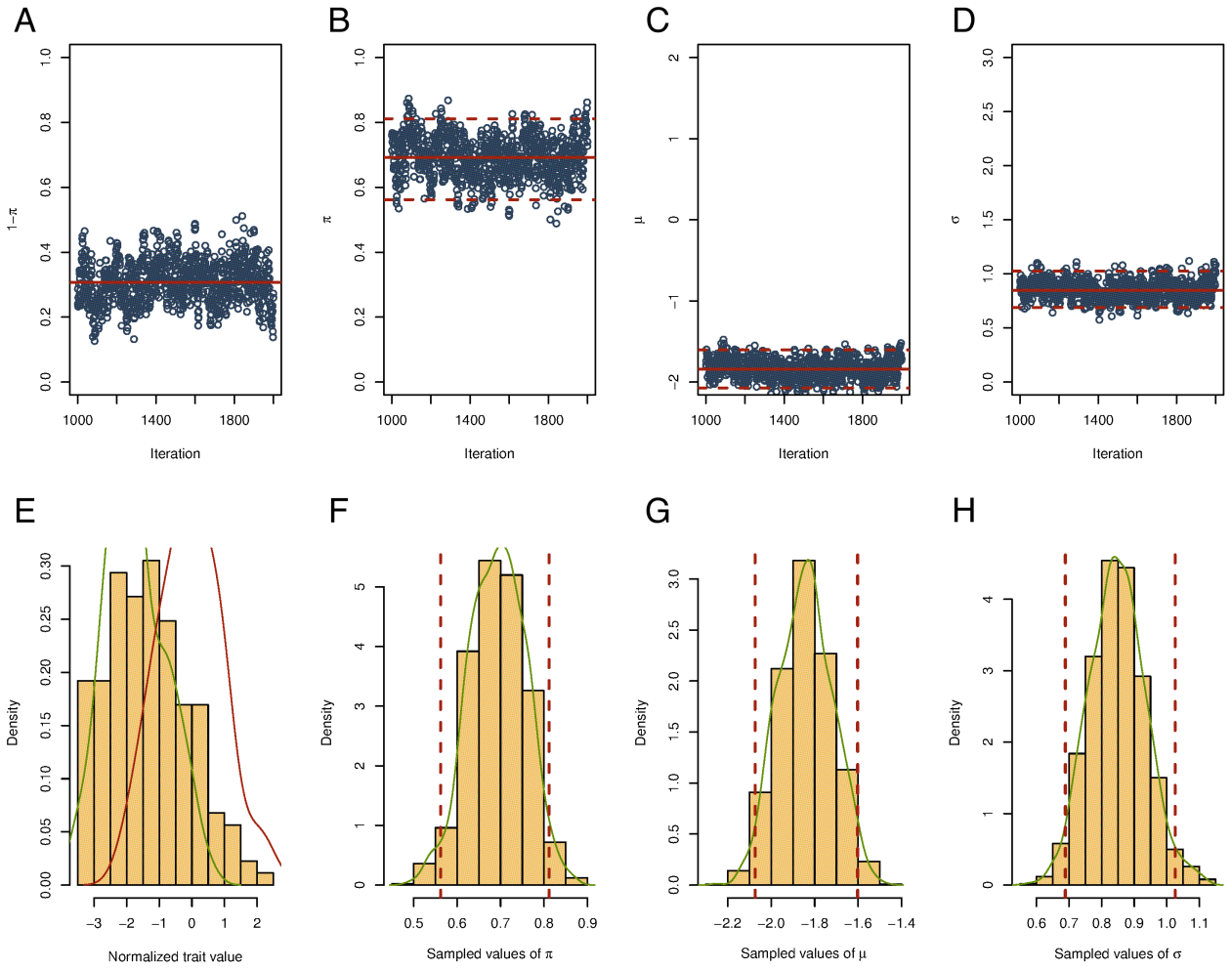
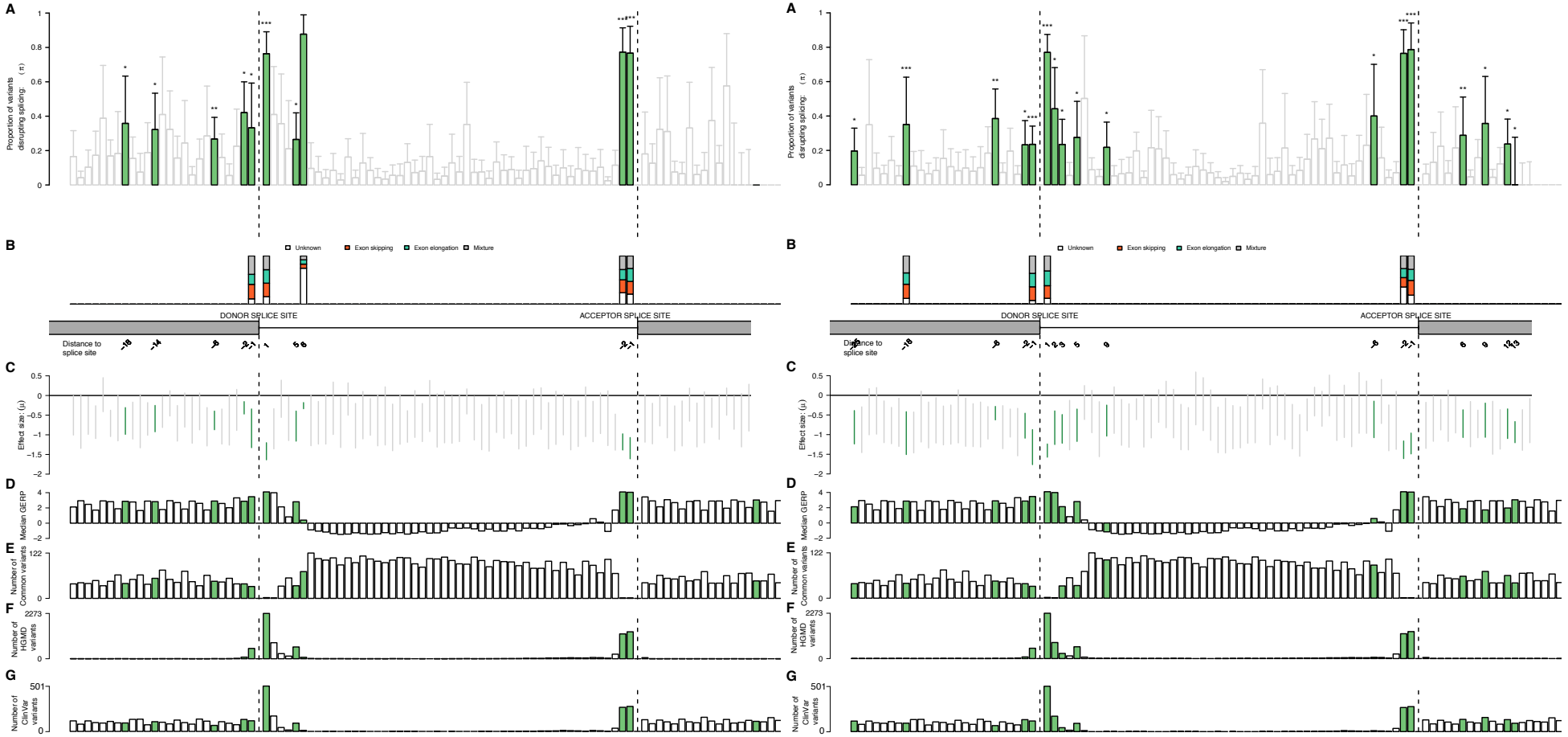
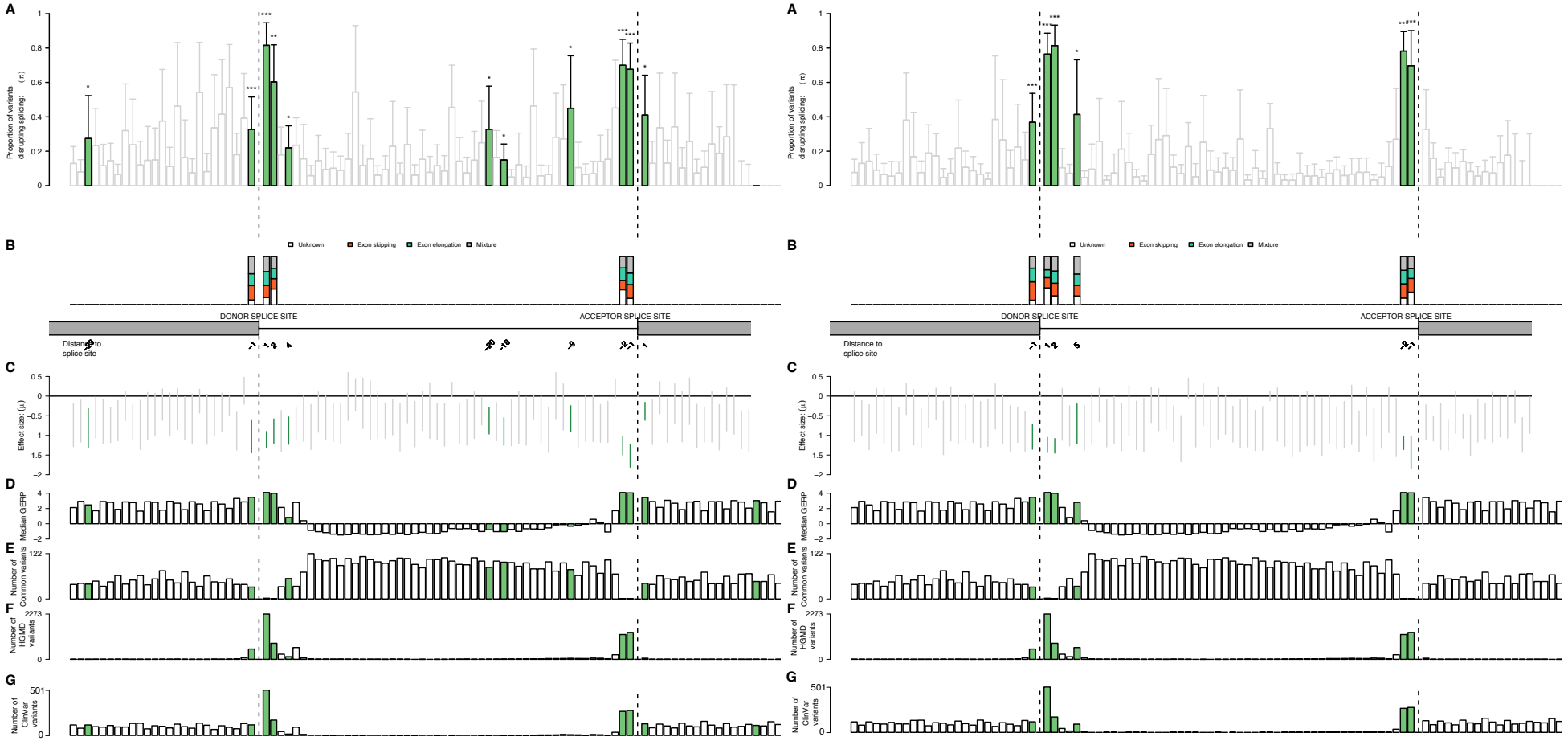


Figure S32: When applying MCMC algorithms to a data set it is customary to show the performance of the algorithm across all stages of the experiment (including the burn-in). We show a typical example of the performance of the SDM algorithm that we use to obtain estimates of the proportion of variants belonging to the group with evidence of splice disruption and estimates of the altrnative shift in the mean μ. We generated these for all distances across all the tissues studied. On the top row we show: A. the estimated proportion of variants belonging to the null splice disruption group (0) across all 2000 iterations; B. the proportion of variants belonging to the splice disruption group (1) across all 2000 iterations, and the mean proportion estimate (solid line) and the 95% confidence interval (dashed line); C. the alternative shift in the mean μ; and D. the variance of the distribution. On the bottom row we show: E. a histogram of the splice quantification values, and the estimated mixture densities; F. a histogram of the estimated proportion of variants belonging to the splice disruption group (1); G. a histogram of the estimated alternative shift in the mean μ; and H. a histogram of the estimated variance.



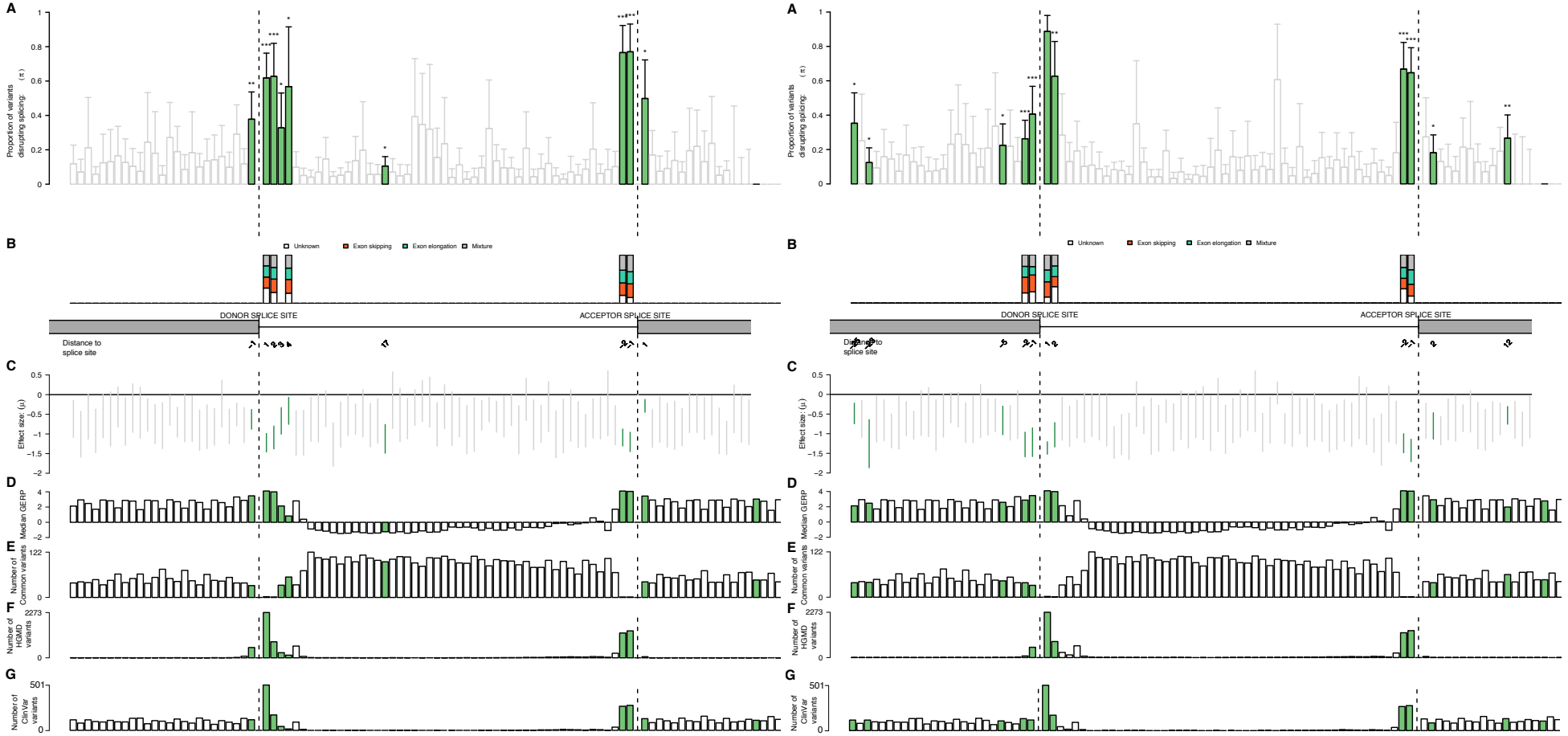
(a) Adipose Subcutaneous **(b)** Artery Tibial

Figure S33: Transcriptional impact of variants proximal to splice junctions: rare variant analysis in Adipose Subcutaneous and Artery Tibial (GTEx data set). A. Proportion of variants disrupting splicing at each distance $\pm$ 1-25bp from donor and acceptor site, ($*$ $P < 0.05$, $**$ $P < 0.01$, $***$ $P < 0.001$; green for $P < 0.05$; SDM p-value evaluated on the estimated proportion of variants supporting the alternative distribution x the effect size of the alternative distribution). B. Classification of splice disruption events: exon skipping (low exon quantification value, no impact on intron quantification), exon elongation (high intron quantification value, no impact on exon quantification), and mixture (high intron and low exon quantification values). C. Effect size estimates (in standard deviations from the population distribution) of the variants on splice junction quantification value. D. Median GERP of all variants and E. Number of common variants identified in an independent exome sequencing study of 4,500 Swedish individuals. F. Number of variants in HGMD version hgmd-2012.4. G. Number of variants in ClinVar (Feb 2015 VCF release) (*55*).



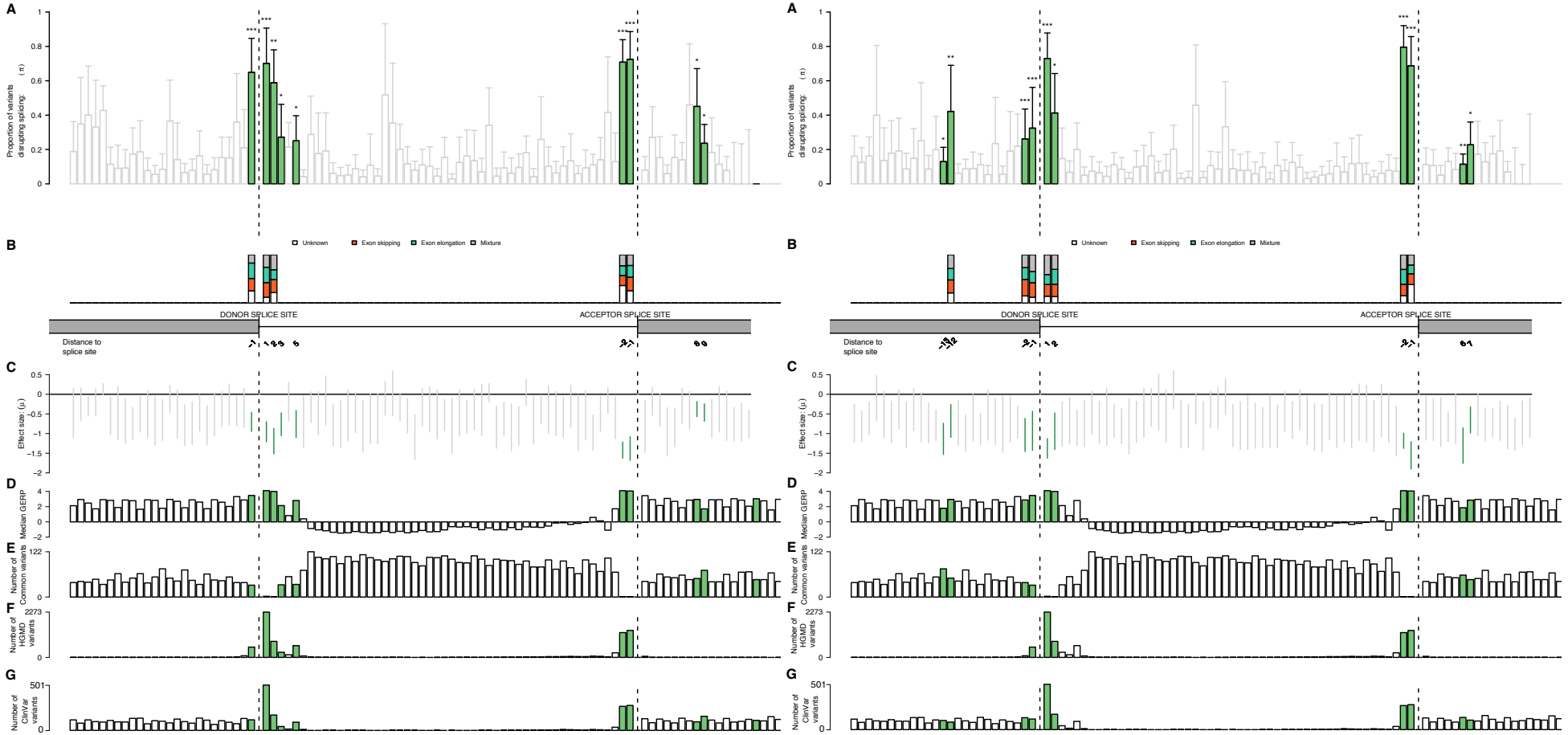
(a) Heart Left Ventricle **(b)** Lung

Figure S34: Transcriptional impact of variants proximal to splice junctions: rare variant analysis in Heart Left Ventricle and Lung (GTEx data set). A. Proportion of variants disrupting splicing at each distance $\pm$ 1-25bp from donor and acceptor site, ($*$ $P < 0.05$, $**$ $P < 0.01$, $***$ $P < 0.001$; green for $P < 0.05$; SDM p-value evaluated on the estimated proportion of variants supporting the alternative distribution x the effect size of the alternative distribution). B. Classification of splice disruption events: exon skipping (low exon quantification value, no impact on intron quantification), exon elongation (high intron quantification value, no impact on exon quantification), and mixture (high intron and low exon quantification values). C. Effect size estimates (in standard deviations from the population distribution) of the variants on splice junction quantification value. D. Median GERP of all variants and E. Number of common variants identified in an independent exome sequencing study of 4,500 Swedish individuals. F. Number of variants in HGMD. G. Number of variants in ClinVar.



(a) Muscle Skeletal **(b)** Nerve Tibial

Figure S35: Transcriptional impact of variants proximal to splice junctions: rare variant analysis in Muscle Skeletal and Nerve Tibial (GTEx data set). A. Proportion of variants disrupting splicing at each distance $\pm$ 1-25bp from donor and acceptor site, ($* P < 0.05$, $** P < 0.01$, $*** P < 0.001$; green for $P < 0.05$; SDM p-value evaluated on the estimated proportion of variants supporting the alternative distribution x the effect size of the alternative distribution). B. Classification of splice disruption events: exon skipping (low exon quantification value, no impact on intron quantification), exon elongation (high intron quantification value, no impact on exon quantification), and mixture (high intron and low exon quantification values). C. Effect size estimates (in standard deviations from the population distribution) of the variants on splice junction quantification value. D. Median GERP of all variants and E. Number of common variants identified in an independent exome sequencing study of 4,500 Swedish individuals. F. Number of variants in HGMD. G. Number of variants in ClinVar.



(a) Skin Sun Exposed **(b)** Thyroid

Figure S36: Transcriptional impact of variants proximal to splice junctions: rare variant analysis in Skin Sun Exposed and Thyroid (GTEx data set). A. Proportion of variants disrupting splicing at each distance $\pm$ 1-25bp from donor and acceptor site, ($* P < 0.05$, $** P < 0.01$, $*** P < 0.001$; green for $P < 0.05$; SDM p-value evaluated on the estimated proportion of variants supporting the alternative distribution x the effect size of the alternative distribution). B. Classification of splice disruption events: exon skipping (low exon quantification value, no impact on intron quantification), exon elongation (high intron quantification value, no impact on exon quantification), and mixture (high intron and low exon quantification values). C. Effect size estimates (in standard deviations from the population distribution) of the variants on splice junction quantification value. D. Median GERP of all variants and E. Number of common variants identified in an independent exome sequencing study of 4,500 Swedish individuals. F. Number of variants in HGMD. G. Number of variants in ClinVar.

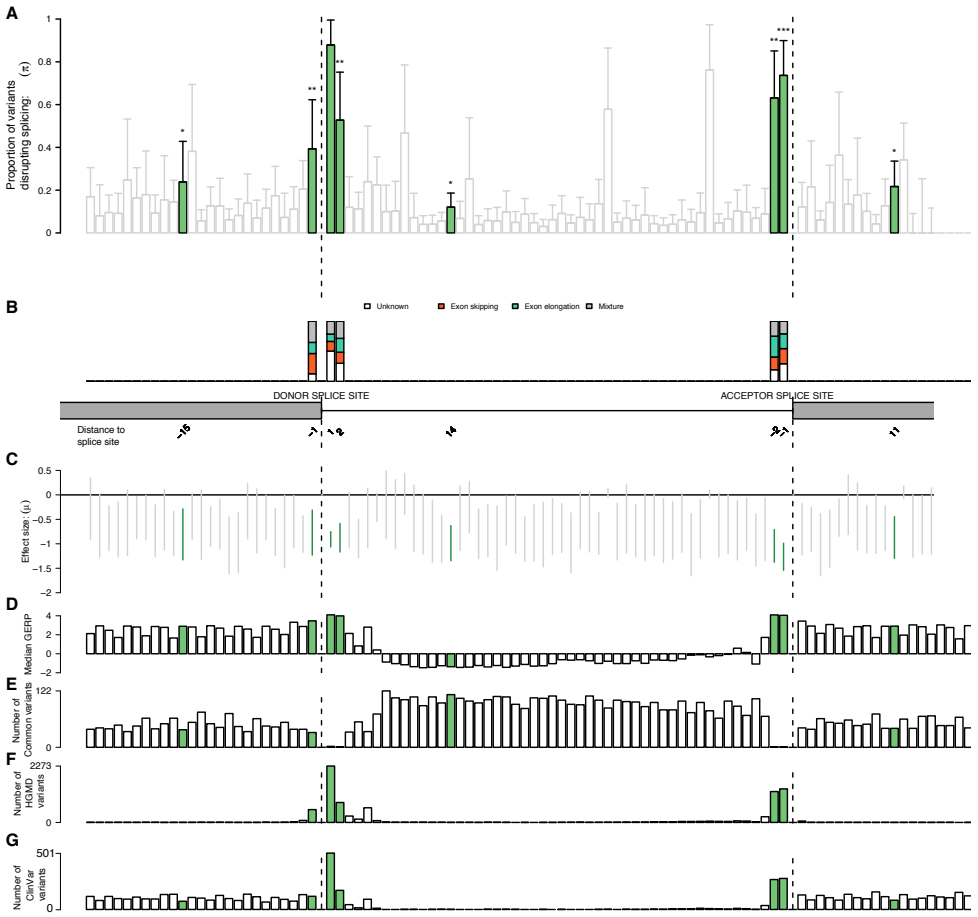


Figure S37: Transcriptional impact of variants proximal to splice junctions: rare variant analysis in Blood (GTEx data set). A. Proportion of variants disrupting splicing at each distance $\pm$ 1-25bp from donor and acceptor site, ($* P < 0.05$, $** P < 0.01$, $*** P < 0.001$; green for $P < 0.05$; SDM p-value evaluated on the estimated proportion of variants supporting the alternative distribution x the effect size of the alternative distribution). B. Classification of splice disruption events: exon skipping (low exon quantification value, no impact on intron quantification), exon elongation (high intron quantification value, no impact on exon quantification), and mixture (high intron and low exon quantification values). C. Effect size estimates (in standard deviations from the population distribution) of the variants on splice junction quantification value. D. Median GERP of all variants and E. Number of common variants identified in an independent exome sequencing study of 4,500 Swedish individuals. F. Number of variants in HGMD. G. Number of variants in ClinVar.

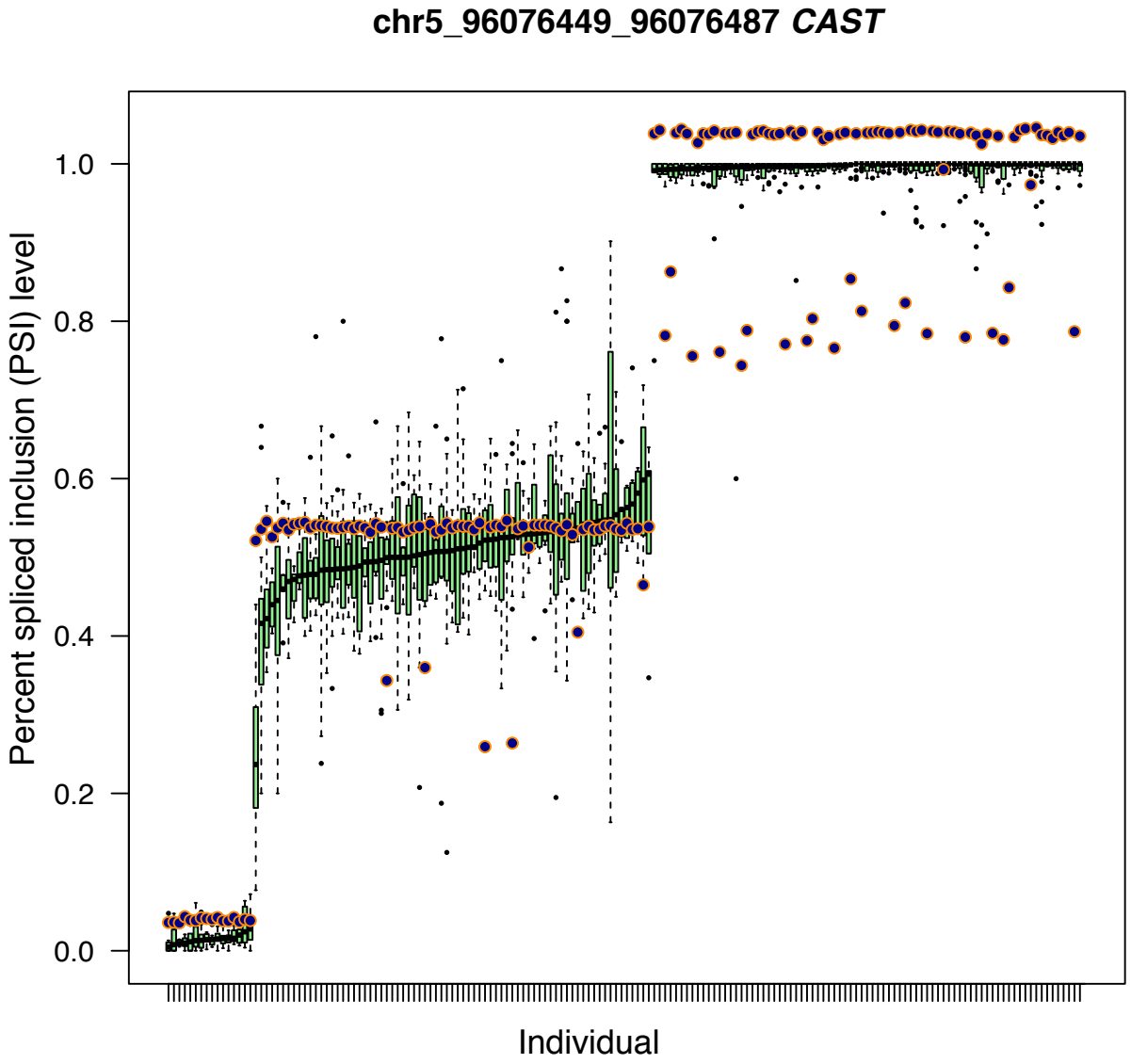


Figure S38: Transcriptional impact of variants proximal to splice junctions: example of a psiQTL variant (rs7724759; $-\log_{10}(P \text{ value}) = 39.07$) for an exon in the gene *CAST* (calpastin). The x-axis represents the 169 individuals included in the analysis. The y-axis represents the exon inclusion level. In the green box plots we show the exon inclusion levels of the tissues for which the individual is sampled; orange dots represent the allelic dosages divided by two so that homozygous reference, heterozygous, and homozygous alternate is approximately 0, 0.5, and 1, respectively. Homozygous individuals either fully exclude (boxplots on the left) or fully include (box plots on the right) the exon, while heterozygous individuals have a partial exon inclusion.

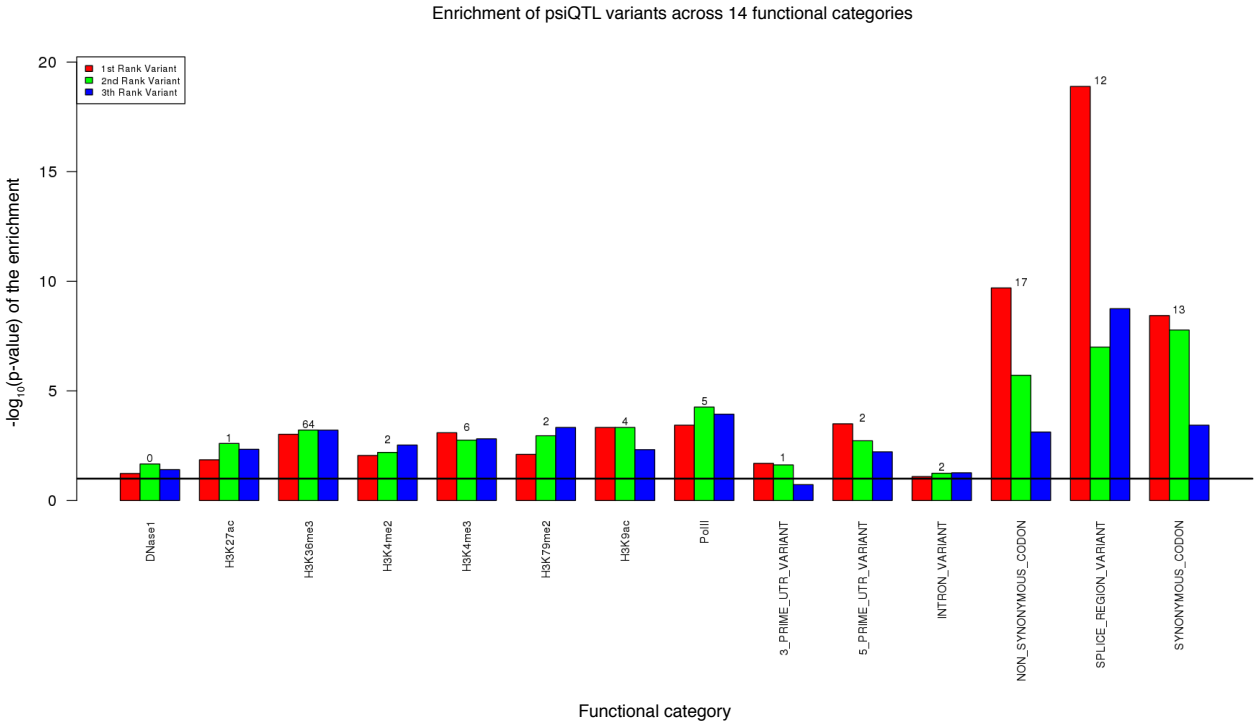
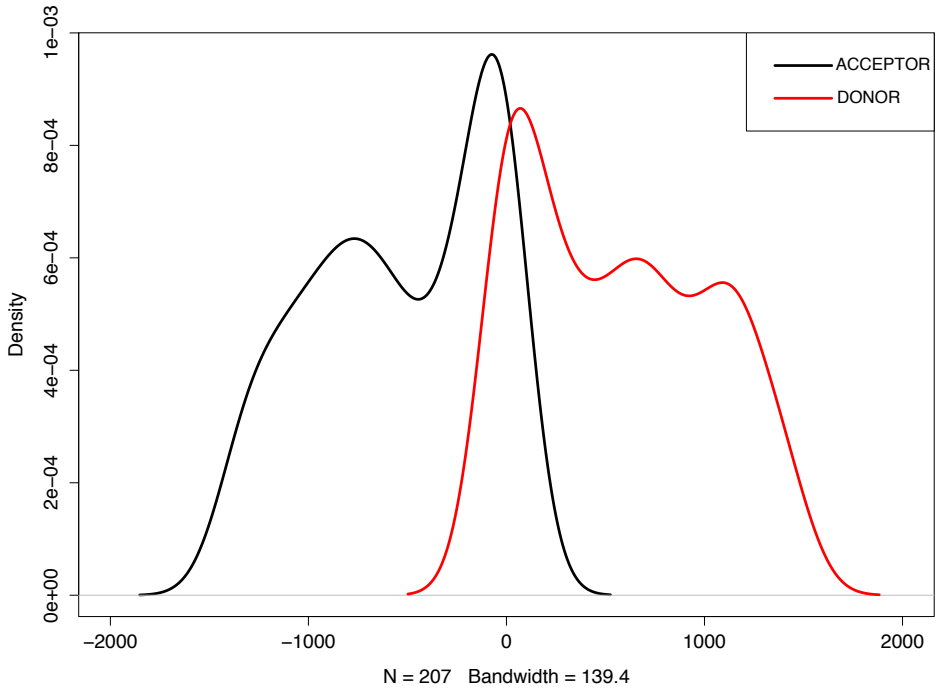
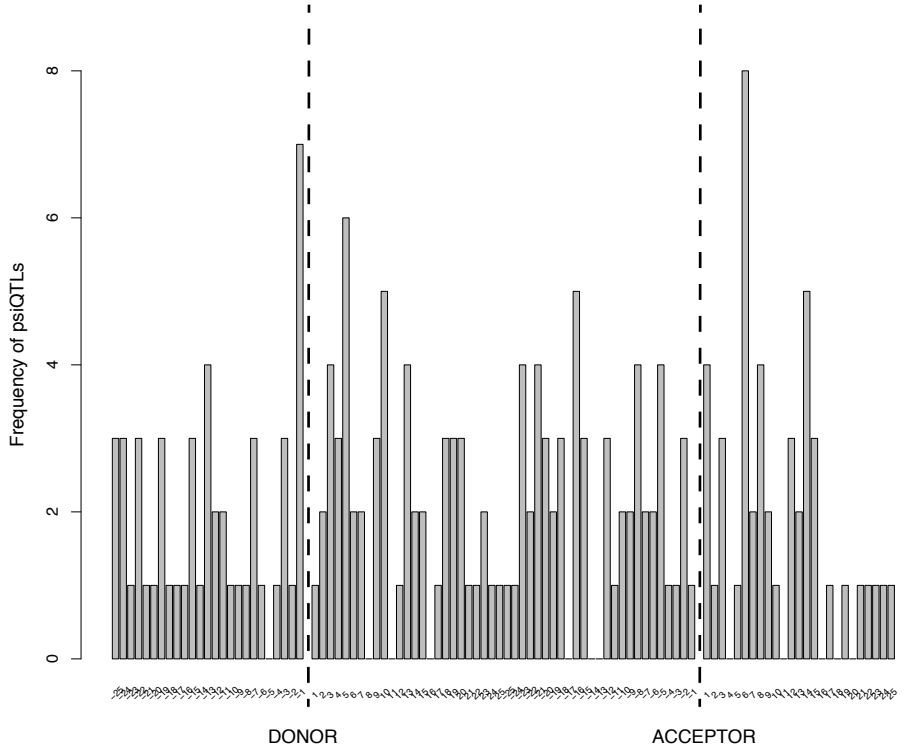


Figure S39: Transcriptional impact of variants proximal to splice junctions: enrichment of common psiQTL variants (for the top 3 associated variants detected in the GTEx data set) across 14 functional categories, calculated as in Lappalainen et al. (*5*). The black line depicts the null of no enrichment. Enrichment of psiQTL variants is observed in splice site regions (SPLICE_REGION_VARIANT).

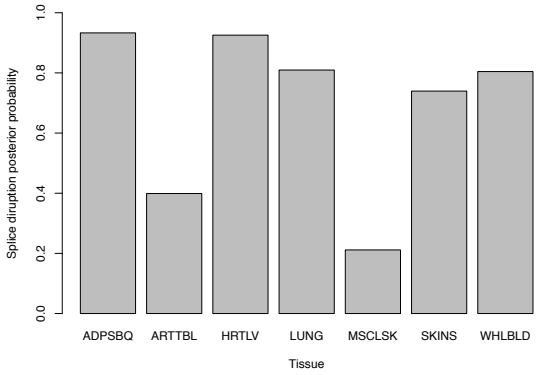


(a) Distance density

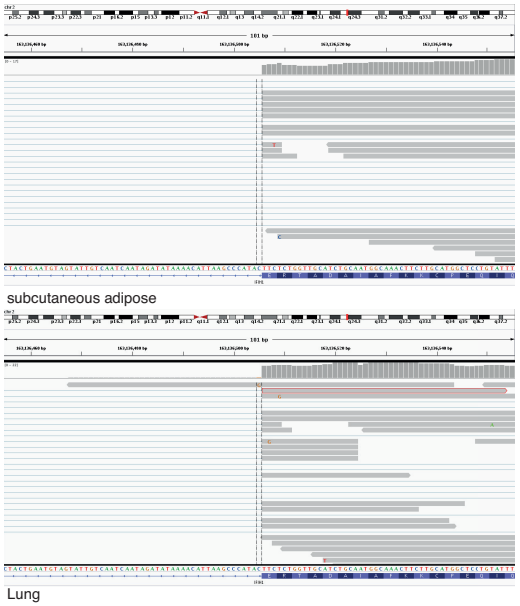


(b) Frequency histogram of psiQTLs proximal to splice junctions

Figure S40: Transcriptional impact of variants proximal to splice junctions: positional patterns of common psiQTL variants. At FDR of 5% ($P = 0.0104$) we find 1779 psiQTLs in 169 tested individuals. We estimate that 1.2% of the exons of a given individual have inclusion levels directly affected by genetic variants which are enriched in splice site regions. a) We show a plot of the density of the distance of variants (for the top associated variant) to the closest splice site with a peak observed between 3 and 50 base pairs (bp). b) The bar plot shows the frequency of common variant psiQTLs proximal to the donor or acceptor junction (50bp window shown).

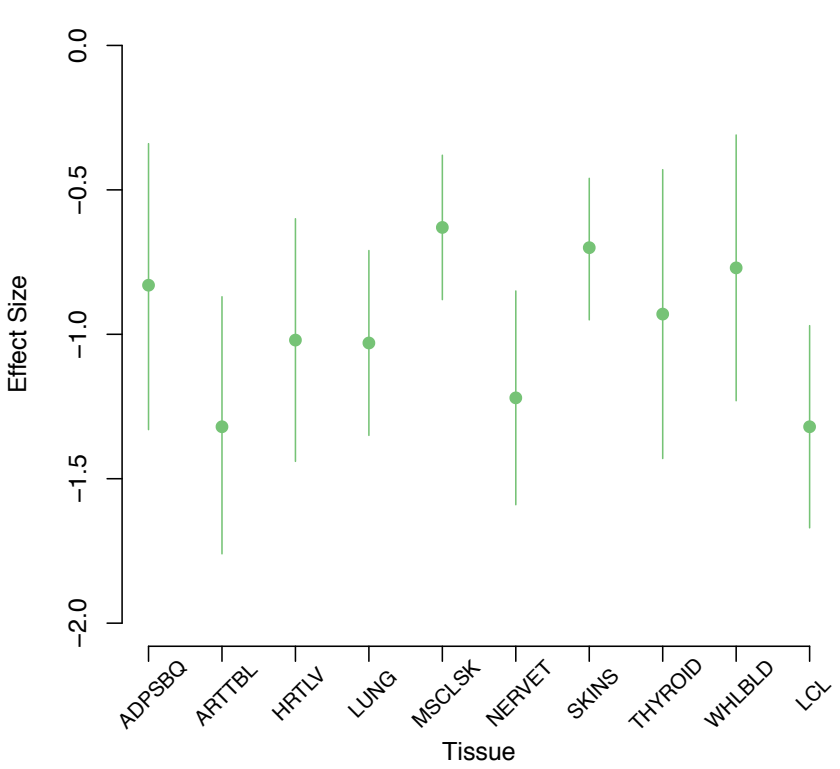


(a) splice disruption posterior probability

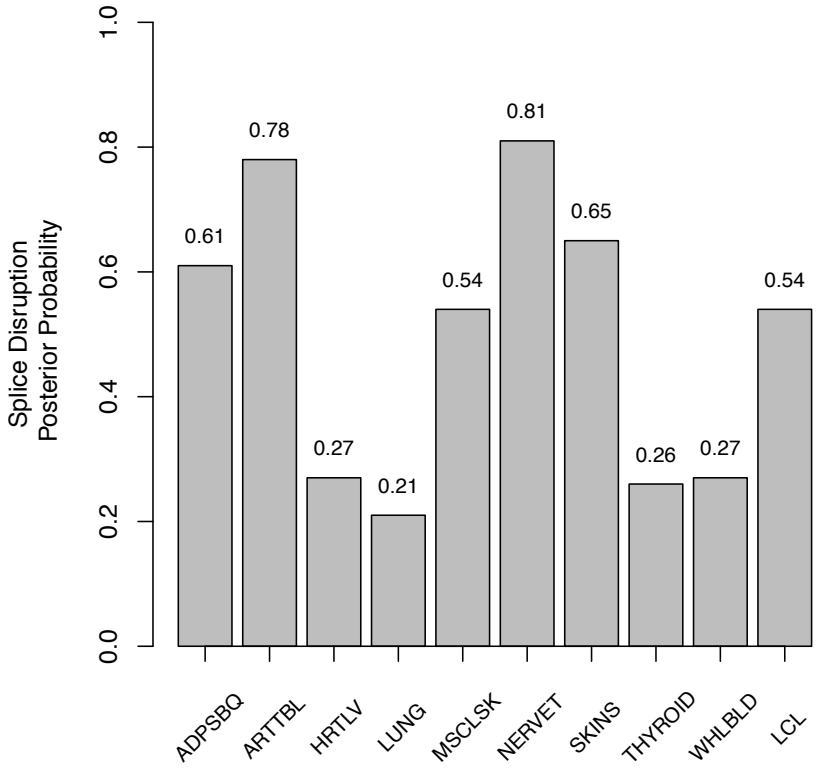


(b) IGV snapshots

Figure S41: Transcriptional impact of variants proximal to splice junctions: a splice disrupting variant, c.IVS8+1G>C (rs35337543), in the gene *IFIH1* (interferon induced with helicase C domain 1). We show a) the splice disruption posterior probability across seven tissues after applying the splice disruption model (SDM) to the GTEx data set and b) IGV snapshot of RNA-seq data from two tissues (subcutaneous adipose and lung) in the c.IVS8+1G>C carrier (top and bottom left). The snapshots show splice disruption in approximately half the reads (light blue lines). We also show IGV snapshot of RNA-seq data from the same two tissues (subcutaneous adipose and lung) in an individual that does not carry the c.IVS8+1G>C PTV variant.



(a) effect size of alternative mixture distribution



(b) splice disruption posterior probability

Figure S42: Transcriptional impact of variants proximal to splice junctions: a splice disrupting variant, rs116928232, in a non-canonical splice site, in the gene *LIPA* (lipase A). We show a) the alternative shift in μ (effect size) for the splice disruption group, and b) the posterior probability that the variant belongs to the splice disruption group across ten tissues in the GTEx data set and Geuvadis data set (LCL). The variant rs116928232 is found in the HGMD data base and annotated as a causal mutation for an autosomal recessive disorder Cholesteryl ester storage disease (CESD, OMIM:27800) associated with reducted activity and genetic defects of lysosomal acid lipase. In (*56*), a compound heterozygote CESD patient is identified. The patient has a nonsense mutation in the paternal allele while the maternal allele contains the A allele in rs116928232. Our data is consistent with the observation that the A allele in rs116928232 disrupts splicing in the patient.

PTV Flags	Description
ANC_ALLELE	PTV is the ancestral allele
NON_CAN_EXON	Exon is surrounded by non-canonical splice site (i.e. not AG/GT)
END_TRUNC	PTV removes less than 5% of remaining protein
SINGLE_EXON	Transcript only has one coding exon
SMALL_INTRON	Splice site variant within intron smaller than 15 bp
NON_CAN_SPLICE	Splice site is non-canonical OR other splice site within same intron is non-canonical
EXON_INTRON_UNDEF	Unable to determine exon/intron boundaries surrounding variant

Table S1: PTV annotation flags used in the annotation pipeline.

Variant type	**GTEx**		**Geuvadis**	
	total (HC)	avg.; homozygous (HC)	total (HC)	avg.; homozygous (HC)
nonsense	1345 (1020)	57.09; 11.56 (29.71; 4.75)	5987 (4682)	71.67; 12.57 (36.55; 3.81)
splice	845 (661)	58.64; 17.17 (29.12; 6.64)	6113 (3252)	125.28; 28.11 (29.39; 4.87)
frameshift	2324 (1746)	107.19; 14.74 (68.10; 6.61)	1023 (606)	16.94; 0.15 (29.73; 0.31)
large deletions	70 (8)	NA	59 (25)	NA
total	4584 (3435)	NA	13182 (8565)	NA

Table S2: Number of PTVs discovered in the GTEx exome sequencing data set and in the Geuvadis/1KG Phase 1 whole-genome data set. Total numbers of PTVs and average number of PTVs per individual ; average number of homozygous PTVs per individual are shown for each PTV class and data set (in parenthesis data shown for variants with HC flags only, i.e. those that do not have any of the filters described in table S1). For the Geuvadis data set we report the numbers for the 421 individuals with genome sequence data. For large deletions we only report the total number of PTV deletion with manual curation in the study.

	Stop-gained HC	Stop-gained LC	Frame-shift HC	Frame-shift LC	Inframe	Synon. in PTV gene	Synon. AF< 0.01	Synon. AF> 0.05	TOTAL
Initial set	151	79	259	119	143	94	50	50	945
Primers designed	102	57	162	78	94	55	40	44	632
Added with longer amp.	20	4	8	2	6	4	5	1	50
Final totals	122	61	170	80	100	59	45	45	682

Table S3: Summary of variant sites selected for mmPCR sequencing experiment in 121 RNA samples from 9 subjects. 682/945 (72%) passed primer design.

Read count	Percentage of PTVs
8	100
15	71
20	57
40	31
60	20
100	9

Table S4: Read depth summary statistic for PTVs studied with ASE data. We compute the percentage of PTVs with a total read count greater than or equal to a read count threshold. Eight is the minimum read count used to include the ASE data in our analyses.

No.	**Predictor**	**Description**
1	X50bp	50bp rule
2	startdist	distance to start codon
3	stopdist	distance to stop codon
4	utr3distend	distance to 3'-UTR end
5	utr5diststart	distance to 5'-UTR start
6	utr3diststart	distance to 3'-UTR start
7	utr5distend	distance to 5'-UTR end
8	utr3size	size of 3'-UTR
9	utr5size	size of 5'-UTR
10	trnscaffected	Variant is annotated as a PTV in all (=FULL) or some (=PARTIAL) transcripts
11	nexons	number of exons
12	ntrnsc	number of alternative isoforms
13	donordist	distance to donor splice site
14	acceptordist	distance to acceptor splice site
15	onecodingexon	indicator variable representing whether the gene contains only one coding exon
16	af	allele frequency
17	GC	Percent GC in a window of +/- 75bp
18	CpG	Percent CpG in a window of +/- 75bp
19	priPhCons	Primate PhastCons conservation score (excl. humans)
20	mamPhCons	Mammalian PhastCons conservation score (excl. humans)
21	verPhCons	Vertebrate PhastCons conservation score (excl. humans)
22	priPhyloP	Primate PhyloP score (excl. humans)
23	mamPhyloP	Mammalian PhyloP (excl. humans)
24	verPhyloP	Vertebrate PhyloP (excl. humans)
25	GerpN	Neutral evolution score defined by GERP++
26	GerpS	Rejected Substitution score defined by GERP++
27	EncExp	Maximum ENCODE expression value
28	EncH3K27Ac	Maximum ENCODE H3K27 acetylation level
29	EncH3K4Me1	Maximum ENCODE H3K4 methylation level
30	EncH3K4Me3	Maximum ENCODE H3K4 trimethylation level
31	EncNucleo	Maximum of ENCODE Nucleosome position track score
32	minDistTSS	Distance to closest Transcribed Sequence Start (TSS)
33	minDistTSE	Distance to closest Transcribed Sequence End (TSE)
34	relcDNApos	Relative position in transcript
35	relCDSpos	Relative position in coding sequence
36	relProtPos	Relative position in protein codon
37	lofflag	LOF flag proposed in MacArthur et al. 2012
38	downstreamexons	Number of exons downstream of the PTV

Table S5: List of 38 predictors used for modeling NMD.

Gene	nptv	ntran	lde	egene	a	d	apval	dpval	rpkm	nzero	tissue
DDT	7	7	MERGED_DEL_2_105343	ENSG00000099977.9	54.46	-45.48	.018	0.24	1803.07	126	LCL
FAM106A	1	1	MERGED_DEL_2_87821	ENSG00000213077.5	8.00	-3.37	4.52×10^{-5}	0.23	14.84	69	LCL
GSTT2	2	2	MERGED_DEL_2_105343	ENSG00000099984.6	1.42	0.88	0.0017	0.25	5.27	126	LCL
LGALS9C	2	2	MERGED_DEL_2_87821	ENSG00000171916.11	24.29	1.85	0.0014	0.87	89.87	69	LCL
OR2T10	1	1	MERGED_DEL_2_8411	ENSG00000184022.2	3.10	2.34	0.13	0.33	12.22	16	LCL
UGT2B17	1	1	CNVR1953.1	ENSG00000197888.2	0.47	-0.014	0.0010	0.94	0.36	17	WHLBLD
UGT2B17	1	1	CNVR1953.1	ENSG00000197888.2	0.70	-0.18	8.17×10^{-06}	0.38	0.66	14	LUNG

Table S6: Linear model results for each gene tested (a $=$ additive effect estimate, d $=$ dominance effect estimate, apval is the *P* value for the Wald test that a $= 0$, dpval is the *P* value that d $= 0$). The addition of a dominance term to the linear model with the additive term never provided a better fit, thus there is no evidence for dosage compensation for these genes in the tissues and donors examined. Gene - HGNC ID, nptv - number of transcripts from this gene annotated as PTV, ntran - total number of transcripts from this gene present in GENCODE, type - "complete" (if nptv $=$ ntran) otherwise partial (all deletions listed in this table are classified as "complete"), lde - large deletion ID for the large deletion overlapping gene, egene - ENSEMBL geneID, RPKM - expression value of deletion carrier, nzero - the number of individuals with copy number 0 at the locus.

Tissue	psiQTLs	variants	exons*
Adipose (ADPSBQ)	417	399	315
Brain	257	253	207
Artery tibial (ARTTBL)	488	468	359
Heart (HRTLV)	351	336	265
Lung (LUNG)	484	464	363
Muscle skeletal (MSCLSK)	511	485	375
Nerve tibial (NERVET)	430	414	332
Skin sun exposed (SKINS)	466	448	352
Thyroid (THYROID)	463	444	351
Whole Blood (WHLBLD)	282	261	209
multiTissue	207	192	144

Table S7: Breakdown of common psiQTL results. $*$Non-overlapping exons.

References and notes

1. J.A. Holbrook, G.Neu-Yilik, M.W.Hentze, A.E.Kulozik, *Nature Genetics* **36**, 801 (2004).
2. P. D. Stenson, et al., *Human Genetics* **133**, 1 (2014).
3. J. C. Cohen, E. Boerwinkle, T. H. Mosley Jr, H. H. Hobbs, *New England Journal of Medicine* **354**, 1264 (2006).
4. D. G. MacArthur, et al., *Science* **335**, 823 (2012).
5. T. Lappalainen, et al., *Nature* **501**, 506 (2013).
6. P. AC't Hoen, et al., *Nature Biotechnology* (2013).
7. J. Lonsdale, et al., *Nature Genetics* **45**, 580 (2013).
8. The GTEx Consortium, *Submitted* (2014).
9. The 1000 Genomes Consortium, *Nature* **491**, 56 (2012).
10. K. R. Kukurba, et al., *PLoS Genetics* **10**, e1004304 (2014).
11. R. Zhang, et al., *Nature Methods* **11**, 51 (2014).
12. S. B. Montgomery, et al., *Genome Research* **23**, 749 (2013).
13. D. A. Skelly, M. Johansson, J. Madeoy, J. Wakefield, J. M. Akey, *Genome Research* **21**, 1728 (2011).
14. K. R. Stevenson, J. D. Coolon, P. J. Wittkopp, *BMC Genomics* **14**, 536 (2013).
15. M. Pirinen, et al., *bioRxiv* (2014). doi:10.1101/007211
16. E. Nagy, L. E. Maquat, *Trends in Biochemical Sciences* **23**, 198 (1998).
17. J. F. Bateman, S. Freddi, G. Nattrass, R. Savarirayan, *Human Molecular Genetics* **12**, 217 (2003).
18. N. Huang, et al., *PLoS Genetics* **6**, 10 (2010).
19. A. A. McAnally, and L. Y. Yampolsky. *Genome Biology and Evolution* **2**, 44-52 (2010).
20. J. Zhou, B. Lemos, E. B. Dopman, D. L. Hartl, *Genome Biology and Evolution* **3**, 1014 (2011).
21. N. A. Faustino, T. A. Cooper, *Genes Development* **17**, 419 (2003).
22. D.J. McCarthy, et al., *Genome Medicine* **6**, 26 (2014).
23. C. B. Burge, T. Tuschl, P. A. Sharp, *Cold Spring Harbor Monograph Archive* **37**, 525 (1999).
24. H. Y. Xiong, et al., *Science* (2014). doi:10.1126/science.1254806
25. A. Corvelo, M. Hallegger, C. W. Smith, E. Eyras, *PLoS Computational Biology* **6**, e1001016 (2010).
26. S. M. Purcell, et al., *Nature* (2014).
27. H. Kacser and J.A. Burns, *Genetics* **97**, 639-666 (1981).
28. M. A. DePristo, et al., *Nature Genetics* **43**, 491 (2011).
29. H. Li, R. Durbin, *Bioinformatics* **25**, 1754 (2009).
30. A. McKenna, et al., *Genome Research* **20**, 1297 (2010).(
31. G. A. Auwera, et al., *Current Protocols in Bioinformatics* pp. 1110 (2013).
32. P. Danecek, et al., *Bioinformatics* **27**, 2156 (2011).(
33. C. Barnes, et al., *Nature Genetics* **40**, 1245 (2008).(
34. M. Fromer, et al., *The American Journal of Human Genetics* **91**, 597 (2012).
35. M. Fromer, S. M. Purcell, *Current Protocols in Human Genetics* pp. 723 (2014).
36. D. F. Conrad, et al., *Nature* **464**, 704 (2010).
37. S. Marco-Sola, M. Sammeth, R. Guigo, P. Ribeca,*Nature methods* **9**, 1185 (2012).
38. O. Stegle, L. Parts, M. Piipari, J. Winn, R. Durbin, *Nature Protocols* **7**, 500 (2012).
39. N. Panousis, M. Gutierrez-Arcelus, E. Dermitzakis, T. Lappalainen, *Genome Biology* (2014).
40. J. Flannick, et al., *Nature Genetics* **46**, 357 (2014).
41. K. Zhang, et al., *Nature Genetics* **38**, 382 (2006).
42. M. Martin, *EMBnet. journal* **17**, pp (2011).
43. A. Dobin, et al., *Bioinformatics* **29**, 15 (2013).
44. C. Trapnell, L. Pachter, S. L. Salzberg, *Bioinformatics* **25**, 1105 (2009).
45. R Development Core Team, *R foundation for Statistical Computing* (2005).

46. R. G. Newcombe, *Statistics in Medicine* **17**, 873 (1998). 75
47. E. B. Wilson, *Journal of the American Statistical Association* **22**, 209 (1927).
48. M. Kircher, et al., *Nature Genetics* **46**, 310 (2014).
49. M. Kuhn, *Journal of Statistical Software* **28**, 1 (2008).
50. X. Robin, et al., *BMC Bioinformatics* **12**, 77 (2011).
51. A. Casadio, et al., *EMBO reports* **16**, 71 (2015).
52. C. N. Henrichsen, E. Chaignat, A. Reymond, *Human molecular genetics* **18**, R1 (2009).
53. C. N. Henrichsen, et al., *Nature genetics* **41**, 424 (2009).
54. A. Schlattl, S. Anders, S. M. Waszak, W. Huber, J. O. Korbel, *Genome research* **21**, 2004 (2011).
55. M. J. Landrum, et al., *Nucleic acids research* p. gkt1113 (2013).
56. C. Aslanidis, et al., *Genomics* **33**, 85 (1996).